Reducing Instability in Synthetic Data Evaluation with a Super-Metric in MalDataGen

Anna Luiza Gomes da Silva[1], **Diego Kreutz**[1], **Angelo Diniz**[1]
Rodrigo Mansilha[1] , **Celso Nobre da Fonseca**[1]

[1]Horizon IA Labs & LEA
Graduate Program in Software Engineering (PPGES)
Federal University of Pampa (UNIPAMPA)

{annaluiza.aluno,diegokreutz,angelonogueira.aluno,rodrigomansilha,celsofonseca}@unipampa.edu.br

Abstract. *Evaluating the quality of synthetic data remains a persistent challenge in the Android malware domain due to instability and the lack of standardization among existing metrics. This work integrates into MalDataGen a Super-Metric that aggregates eight metrics across four fidelity dimensions, producing a single weighted score. Experiments involving ten generative models and five balanced datasets demonstrate that the Super-Metric is more stable and consistent than traditional metrics, exhibiting stronger correlations with the actual performance of classifiers.*

1. Introduction

Synthetic data generation has become an increasingly relevant strategy in cybersecurity [Figueira and Vaz 2022, Lee 2025, Hao et al. 2024], particularly as a way to mitigate the scarcity of real, complete, and high-quality datasets that limit the performance and generalization of machine learning models. Despite these advances, assessing the quality of synthetic data remains a complex and largely non-standardized methodological challenge [Platzer and Reutterer 2021], with no clear consensus on which metrics should be used or how to combine them consistently.

The literature reports a significant fragmentation in the application of fidelity metrics, with studies identifying more than 65 distinct indicators used independently to assess fidelity [Silva et al. 2025]. This diversity hinders model-to-model comparison, reduces experimental reproducibility, and complicates the integrated interpretation of data quality. Tools such as the Synthetic Data Vault (SDV)[1], which implements Copula, TVAE, and CTGAN [Patki et al. 2016]; YData Synthetic[2], which offers multiple variations of GANs; and Gretel Synthetics[3], which uses models such as DGAN, DPGAN, and ACTGAN, attempt to consolidate generation and evaluation processes. Additionally, initiatives such as [Dahmen and Cook 2019] demonstrate the application of HMMs for time-series generation in the healthcare domain. However, these platforms exhibit limitations related to flexibility, restricted customization capabilities, and relatively small sets of pre-implemented algorithms.

To address these limitations, this work[4] extends the MalDataGen framework [Paim et al. 2025], a modular open-source platform for generating synthetic tabular data,

[1]https://sdv.dev
[2]https://ydata.ai/
[3]https://gretel.ai/
[4]Pre-print version available on arXiv [da Silva et al. 2025].

through the integration of a generalizable Super-Metric developed to unify fidelity assessment [Silva et al. 2025]. The Super-Metric combines eight metrics organized into four foundational dimensions, namely Distance, Correlation and Association, Feature Similarity, and Multivariate Distribution, and produces a single weighted score that reduces the variability and inconsistency typically observed when metrics are applied independently.

The main contribution of this work is the evolution of MalDataGen from a data generation tool into a comprehensive ecosystem for multidimensional generation and evaluation of synthetic datasets aimed at Android malware detection. With the integration of the Super-Metric, the framework delivers a more robust, stable, and context-aware evaluation process, improving its suitability for critical cybersecurity applications.

2. Related Work

Evaluating the fidelity and utility of synthetic data remains a challenging and fragmented task. Existing utility metrics frequently lead to conflicting conclusions, which complicates reliable comparison of synthetic data generators (SDGs), particularly in structured or high-dimensional domains where sparsity and complex dependencies amplify inconsistencies. To bring structure to this problem, [Dankar et al. 2022] propose a multidimensional framework that groups metrics into four categories: attribute fidelity, bivariate fidelity, population fidelity, and application fidelity. Their results show that metrics commonly disagree on which generator performs best because each captures different statistical properties, leading to contradictory assessments.

A complementary line of work investigates whether broad, model-level metrics can predict task-specific utility. In a large-scale study with 30 health datasets, [El Emam et al. 2022] demonstrate that most multivariate measures fail to consistently reflect downstream predictive performance. Only the multivariate Hellinger distance exhibited reliable behavior, while metrics such as Maximum Mean Discrepancy, Wasserstein distance, clustering measures, and distinguishability tests showed weak or unstable predictive ability. The study also highlights substantial variability introduced by SDG stochasticity, underscoring the importance of stability as an essential but often overlooked dimension of evaluation.

These findings reveal key gaps in the literature: the lack of an integrated metric capable of aggregating heterogeneous fidelity indicators, the absence of mechanisms to adapt or learn metric weights according to downstream utility, and persistent instability across both broad and narrow evaluation measures. The Super-Metric introduced in this work addresses these limitations by combining multiple fidelity metrics across four core dimensions, learning optimal weights to align with utility metrics such as recall and F1-score, and reducing variance across datasets and generators. As a result, it advances synthetic data evaluation toward greater stability, reproducibility, and practical relevance.

3. The MalDataGen Framework

MalDataGen [Paim et al. 2025, Nogueira et al. 2025] is a modular and open-source framework designed to systematically and reproducibly orchestrate the generation and evaluation of synthetic tabular data in the context of Android malware detection. Its goal is to provide a unified platform that enables the comparison of different generative models

under the same experimental methodology, reducing implementation bias and ensuring consistency across executions.

The framework's architecture is organized into three main components: (i) *input and preprocessing*, responsible for standardizing *datasets*, normalizing attributes, and preparing data for the generative models; (ii) *generative layer*, which integrates multiple families of models capable of synthesizing tabular datasets with varying levels of complexity; and (iii) *evaluation layer*, which computes traditional fidelity and utility metrics, as well as the Super-Metric integrated in this work. Figure 1 presents a diagram illustrating the workflow across these components.

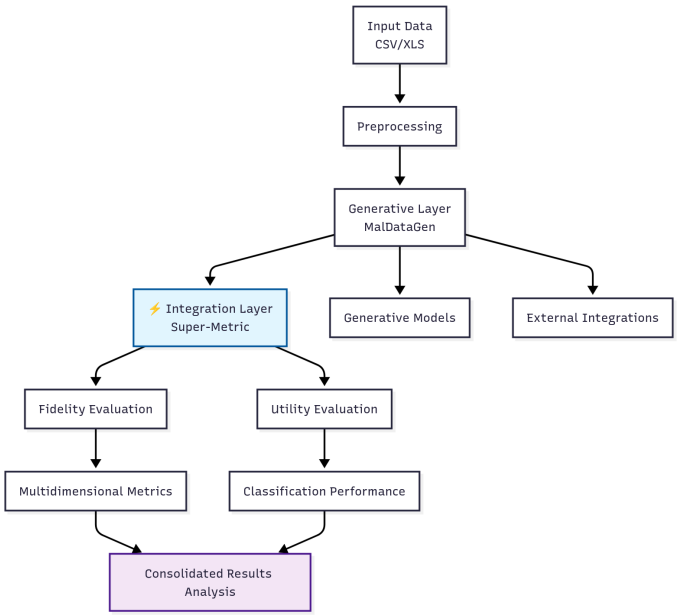

Figura 1. Workflow of the synthetic data generation and evaluation methodology in MalDataGen.

A central pillar of MalDataGen is its function as a flexible benchmark, enabling different generation paradigms to be evaluated under identical conditions. To support this, the generative layer includes four groups of models: (i) *Adversarial Models* (GANs): classical GAN, WGAN, and WGAN-GP; (ii) *Autoencoders*: standard autoencoder, VAE, and quantized VAE; (iii) *Diffusion Models*: Denoising Diffusion and Latent Diffusion; (iv) *Statistical and third-party models*: SMOTE and SDV library models (CTGAN, TVAE, Copula).

The metrics module in MalDataGen organizes the evaluation process into three categories: binary metrics (e.g., precision, recall, F1-score), distance metrics (such as Euclidean and Hellinger), and probabilistic metrics (such as AUC-ROC). The internal infrastructure standardizes result storage by evaluation strategy, classifier, and fold, ensuring traceability and comparability across experiments.

The Super-Metric, integrated as a composite metric within the distance module, extends the evaluation system by providing a consolidated multidimensional analysis. It combines eight metrics distributed across four fundamental dimensions: distance, association, feature similarity, and multivariate distribution, and produces a single weighted

final score. Its integration occurs transparently within the framework's internal workflow, using the same routines and data structures as conventional metrics.

With this integration, MalDataGen evolves from a data generation tool into a complete ecosystem for generating, evaluating, and rigorously comparing generative models applied to the Android malware domain. This enables more consistent, stable, and comparable analyses, contributing to reproducible and methodologically sound experiments in cybersecurity.

4. The Super-Metric

Evaluating the quality of synthetic data remains a central challenge in the generation of tabular datasets, as traditional fidelity and utility metrics, such as statistical distances, distribution divergences, and association measures, tend to capture only specific aspects of the problem. These metrics are generally domain-sensitive, exhibit instability in the presence of multimodal distributions, and can be difficult to interpret collectively, especially in scenarios characterized by strong class imbalance and highly sparse binary attributes, as commonly observed in Android malware data. Although composite approaches such as TabSynDex [Chundawat et al. 2024] represent advances by consolidating multiple dimensions into a single index, previous studies [Silva et al. 2025] indicate that they may still present significant variation across generative models and do not always reflect the real impact of synthetic data on the performance of supervised classifiers.

In this context, the Super-Metric was designed to provide a more robust and informative alternative. Its formulation combines different fidelity dimensions in a weighted manner, where the weights are not learned during synthetic data generation, but are computed afterward, once the utility metrics (such as *recall* and F1-score) have been obtained. These weights are optimized post hoc to maximize the correlation between the aggregated fidelity score and the actual downstream performance of classifiers. Rather than relying on uniform aggregation, the Super-Metric operates as an optimized composition that highlights which fidelity dimensions are most aligned with utility, capturing both structural similarity and discriminative patterns relevant to malware detection. In this way, it functions as a global quality indicator and as a predictive estimator of how synthetic data is expected to perform in real classification scenarios.

5. Evaluation

To assess the effectiveness of the Super-Metric integrated into MalDataGen, we conducted experiments involving ten generative models and five balanced Android malware *datasets* [Nawshin et al. 2024, Alomari et al. 2023]. For each combination of *dataset* and generator, we computed traditional fidelity metrics as well as the proposed Super-Metric, comparing them against utility metrics (recall and F1-score) obtained from classifiers trained on synthetic data and evaluated exclusively on real data. It is important to note that some fidelity metrics appear more than once because MalDataGen adopts distance-based implementations, while the Super-Metric uses similarity-based variants of the same measures. For instance, Jaccard, Hellinger, and Hamming are computed as distances in MalDataGen, but as similarities within the Super-Metric. Although conceptually related, these formulations behave differently, distance metrics penalize divergence, whereas similarity metrics reward alignment. Therefore, both versions are intentionally preser-

ved in the evaluation, and their appearance as "duplicates" reflects distinct computational definitions rather than redundancy.

In all experiments, the Super-Metric was computed independently for each dataset, with the objective of reducing the gap between recall and F1-score while preserving statistical fidelity. This strategy ensures that the final score captures both structural similarity and the practical utility of synthetic data in real classification tasks.

The analysis relied on two main types of visualizations: heatmaps representing the average correlation between each fidelity metric and the utility metrics, and boxplots depicting the distribution of these correlations across different generators. These visualizations supported the assessment of three desirable properties for fidelity evaluation: consistency, defined as preserving the same correlation sign; stability, associated with low variance across generative models; and robustness, expressed as behavior independent of the chosen generator. Positive correlations indicate that fidelity improvements align with enhanced classification performance. Metrics with negative or near-zero correlations offer limited predictive value for practical utility. Figures 2 and 3 summarize these findings. Figure 2 presents the average correlations organized by generator model, while Figure 3 shows the distribution of correlations across datasets. Together, the results highlight the advantages of the Super-Metric in heterogeneous generation scenarios.

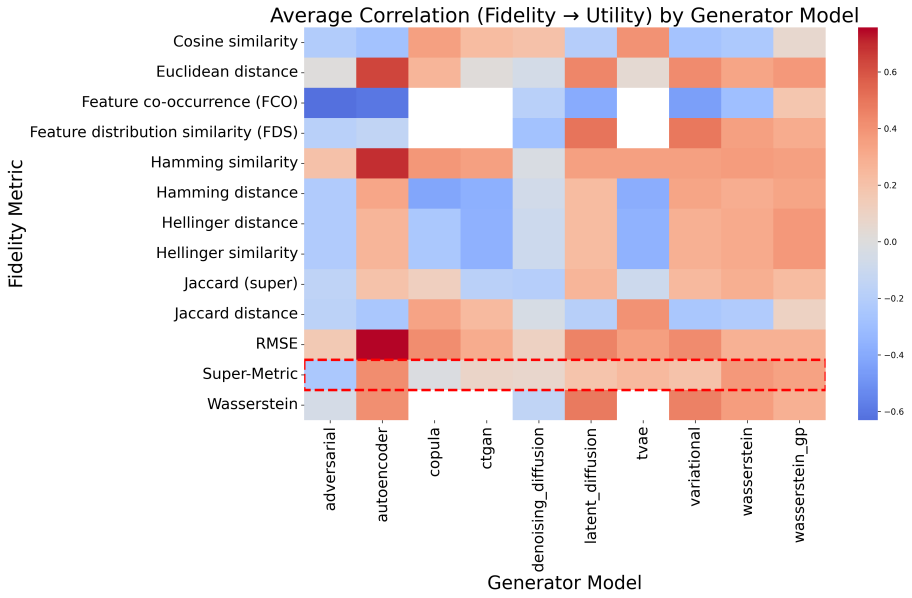

Figura 2. Heatmap – Average correlation between fidelity metrics and utility metrics (recall and F1-score) per generative model.

The results show that traditional metrics exhibit highly unstable behavior, alternating between positive and negative correlations and displaying large variation across generative models. This behavior indicates that none of these metrics can serve as a universal fidelity metric. In contrast, the Super-Metric demonstrates greater stability, a consistent correlation sign, and better alignment with recall and F1-score. Even when it does not achieve the highest absolute correlation, it stands out as the most stable metric across generators, indicating that its weighted aggregation reduces noise and mitigates limitations present in metrics evaluated in isolation.

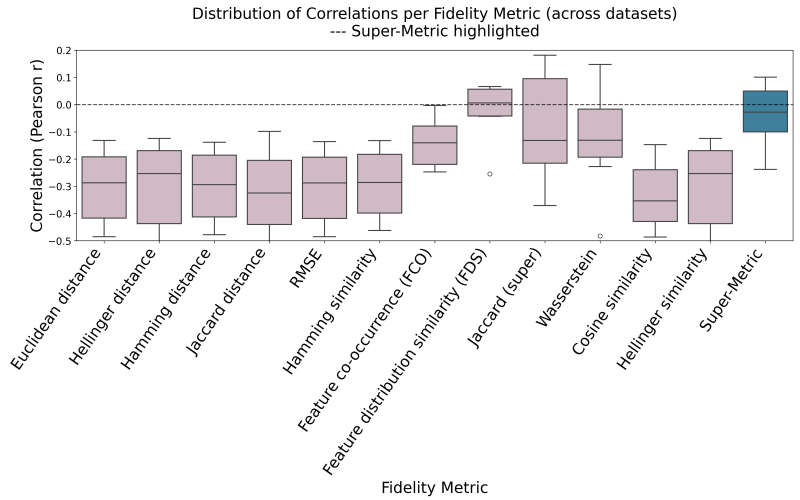

Figura 3. Boxplot – Distribution of the correlation between fidelity metrics and utility metrics (recall and F1-score) across datasets.

The Super-Metric stands out not only for the magnitude of its correlations, but primarily for its stability across generator models and consistency of positive sign. While individual metrics like FDS may occasionally achieve higher correlations in specific scenarios, they frequently exhibit significant variations between models and datasets. Specifically, FDS shows null correlations for three generator models and negative correlations for others, despite its low inter-dataset variation. In contrast, the Super-Metric maintains reliable positive correlations across all evaluation dimensions.

6. Final Considerations and Future Work

This work integrated a fidelity-oriented Super-Metric into MalDataGen, extending the framework beyond synthetic data generation and establishing it as a robust platform for benchmarking. The inclusion of diverse generative models, such as GANs, autoencoders, statistical approaches, and diffusion models, enabled a comprehensive comparative analysis across heterogeneous generation scenarios. The experimental results showed that traditional fidelity metrics tend to produce inconsistent behaviors and exhibit high sensitivity to the underlying generator, which leads to low stability and fluctuating correlations with utility metrics. In contrast, the Super-Metric presented more consistent performance, lower variance across models, and stronger alignment with the actual behavior of classifiers, addressing a significant gap in the evaluation of synthetic data quality in the malware domain.

Future work includes improving the Super-Metric through advanced optimization techniques to enhance its discriminative capabilities, extending the approach to other domains, and incorporating interpretability mechanisms to better understand the contribution of each metric dimension. Another promising direction is its integration into MLOps pipelines to enable continuous monitoring of synthetic data quality in real operational environments, further strengthening its role within the MalDataGen ecosystem.

Acknowledgments. This research was partially supported by CAPES[5], under Financing

[5]https://www.gov.br/capes/pt-br

Code 001, and by FAPERGS[6], through grant agreements 24/2551-0001368-7, 24/2551-0000726-1 and 22/2551-0000841-0.

Referências

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).

Chundawat, V. S., Tarun, A. K., Mandal, M., Lahoti, M., and Narang, P. (2024). Tabsyndex: A universal metric for robust evaluation of synthetic tabular data.

da Silva, A. L. G., Kreutz, D., Diniz, A., Mansilha, R., and da Fonseca, C. N. (2025). Reducing instability in synthetic data evaluation with a super-metric in MalDataGen.

Dahmen, J. and Cook, D. (2019). Synsys: A synthetic data generation system for healthcare applications. *Sensors*, 19(5).

Dankar, F. K., Ibrahim, M. K., and Ismail, L. (2022). A multi-dimensional evaluation of synthetic data generators. *IEEE Access*, 10:11147–11158.

El Emam, K., Mosquera, L., Fang, X., and El-Hussuna, A. (2022). Utility metrics for evaluating synthetic health data generation methods: Validation study. *JMIR Med Inform*, 10(4):e35734.

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

Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., Zhou, Z., and Tang, H. (2024). Synthetic data in ai: Challenges, applications, and ethical implications. *arXiv preprint arXiv:2401.01629*.

Lee, P. (2025). Synthetic data and the future of ai. *Cornell L. Rev.*, 110:1.

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.

Nogueira, A. G. D., Paim, K. O., Bragança, H., Mansilha, R. B., and Kreutz, D. (2025). Synthetic data: Ai's new weapon against android malware.

Paim, K., Nogueira, A., Kreutz, D., Cordeiro, W., and Mansilha, R. (2025). MalDataGen: A modular framework for synthetic tabular data generation in malware detection. In *Anais Estendidos do XXV SBSeg*, Porto Alegre, RS, Brasil. SBC.

Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The synthetic data vault. In *IEEE DSAA*.

Platzer, M. and Reutterer, T. (2021). Holdout-based empirical assessment of mixed-type synthetic data. *Frontier in Big Data*.

Silva, A., Nogueira, A., Kreutz, D., Paim, K., Mansilha, R., and Fonseca, C. (2025). Além da similaridade: Uma super-métrica generalizável para avaliação de fidelidade em dados sintéticos de malware. In *Anais do XXV SBSeg*. SBC.

[6] `https://fapergs.rs.gov.br`