

Curating the Collapse: How Sample-Level Auditing Accelerates Model Collapse in Synthesis of Tabular Data

William Martin Stevenson¹, Anne Gerdes², Lukas Galke Poech¹, Anton Danholt Lautrup^{1*}

¹ Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

² Department of Design, Media and Educational Science, University of Southern Denmark, Odense, Denmark
wiste24@student.sdu.dk, gerdes@sdu.dk, galke@imada.sdu.dk, lautrup@imada.sdu.dk

Abstract

Generative artificial intelligence has fundamentally changed our relationship with data, and the proliferation of artificially generated, *synthetic*, content on the internet has sparked much concern from the community. Because it is difficult to moderate and identify high-fidelity synthetic examples, future deep learning algorithms and AI systems based on data sourced from the internet are at risk of reciting spurious patterns, defects, and unfairness. Moreover, unchecked synthetic data reuse in generative models drives a self-replicating cycle of deterioration, often under the umbrella term *model collapse*. In this work, we focus on the synthesis of tabular data, which has received broad attention in high-risk application domains such as healthcare and finance, for dataset augmentation and privacy preservation, but remains understudied for model collapse dynamics. We demonstrate how the effect persists for tabular data and extend our findings to show how various levels of sample curation (output quality filtering) can provoke mode dropping and diversity loss. Finally, we discuss trade-offs associated with auditing and propose future directions for synthetic data generation pipelines.

Code — <https://github.com/wmstevenson/CuratedCollapse>

Introduction

Recent advances in generative artificial intelligence have made synthetic data of all types readily available for anyone to create and use for better or worse. Artificially generated synthetic content now makes up a volume of the total data in the world that many consider alarming (Schaeffer et al. 2025). Most notably, the wide proliferation of AI-generated misinformation content, bots, and deepfakes across social media platforms has raised concerns for the long-term health of these platforms (Lundberg and Mozelius 2025; Park and Nan 2026). These concerns are not limited to the direct harms of misleading or fabricated content. They also raise broader questions about provenance and authenticity in increasingly automated information ecosystems, where synthetic outputs may circulate and become detached from their original context (Chandra et al. 2024).

A related issue is raised for the future of learning algorithms: training data sourced from the internet may be

contaminated or largely made up of synthetic outputs of their predecessors (Martínez et al. 2024). This recursive reuse of synthetic data can result in model collapse, in which inaccuracies and defects in the generative model or its synthetic outputs compound across generations (Shumailov et al. 2024; Alemohammad et al. 2024; Koul, Duran, and Hernandez-Boussard 2025; Cloud et al. 2026).

At the same time, purpose-built synthetic data is increasingly offered as a solution for scaling data in domains where privacy protection, limited data availability, and collection cost are restrictive (Hernandez et al. 2022; Bieringer et al. 2022; Lautrup et al. 2024). Dataset augmentation using generative models has shown success in generating synthetic data that can supplement and/or diversify the existing training data such that downstream model performance is improved (Frid-Adar et al. 2018; Chen, Mowlaei, and Shi 2020). This creates a tension between the practical issues that synthetic data solves and the technical risks associated with downstream use.

This work focuses on structured tabular data, which underpins many institutional decision systems across domains such as health, finance, education, and public administration. Relative to other data modalities (image, text), tabular data is under-represented in discussions of synthetic content and model collapse across the literature. While tabular data may, at first glance, seem less vulnerable than other modalities, practices such as widespread sharing, re-sharing, open-sourcing of datasets and dataset variants, with imputations and synthetic augmentations, can still obscure traceability to an authentic data source.

Of particular concern are cases where the datasets are already synthetic, for example, created to be shared alongside a medical study as a proxy for the real sensitive patient records (Bhanot et al. 2021). Moreover, as data science becomes increasingly automated with Agentic AI assistants conducting analyses and producing results, the likelihood of unknowing synthetic data reuse could be even greater. Finally, while we have become accustomed to looking for artifacts and hallucinations in images and language model outputs, noticing subtle distributional drifts in tabular data may be much more challenging. Such shifts may have serious consequences for algorithmic fairness long before the resulting biases become pronounced enough to be obvious (Wyllie, Shumailov, and Papernot 2024; Koul, Duran,

*Corresponding Author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

and Hernandez-Boussard 2025).

Contributions. This paper demonstrates model collapse dynamics for tabular data and aims to initiate a discussion of model collapse in relation to tabular data, an imperative if our institutions expect continued reliance on tabular insights. Specifically, we make the following contributions:

- We demonstrate that generative models for tabular data are at risk of similar collapse dynamics to their better-researched image and language model counterparts.
- We show how post hoc model auditing (filtering for high-quality samples) can influence model collapse dynamics.
- We recommend future directions for working with synthetic-data pipelines and how to track degradation.

Background and Related Work

Model collapse is a recent concept that warns about risks involved with the indiscriminate use of artificially generated content in model training. The term *model collapse* is often attributed to Shumailov et al. (2024), who popularised its idea, but earlier works such as (Shumailov et al. 2023) and (Alemohammad et al. 2024) also cover these phenomena. Importantly, many empirical and theoretical results of these papers were concluded to be irrespective of model architecture and data type.

At a high level, model collapse refers to a degenerative process across instances of generative model outputs. Synthetic examples from previous iterations of models ‘pollute’ the training set of subsequent generations, causing the accumulation of error types. As a result, downstream models and systems may learn a distorted perception of reality. The concept is intuitive to understand; however, at the time of writing, there is little consensus on formal definitions of the phenomenon. Schaeffer et al. (2025) exposed the extent to which definitions diverged across the existing literature. The authors reviewed 28 contemporary works and distinguished eight different definitions of model collapse, which are used haphazardly across the literature, indeed also within defining works such as (Shumailov et al. 2023).

In the following, we will refer to model collapse broadly as observable degradation across one or more evaluation metrics. This gives a broad definition under which we can speculate which error sources resulted in a specific type of collapse. The most prominent error sources (Shumailov et al. 2024) that can compound over generations and cause deviation from the original model are as follows,

- **Statistical Approximation Error:** Larger datasets allow modelling of the underlying distribution with higher precision. However, because real-world datasets contain a finite (and often low) number of samples, errors can arise from random information loss due to samples that are not carried through in successive steps of resampling.
- **Functional Expressivity Error:** The capacity of the generative model can also lead to errors, particularly if the capacity is low. Namely, the model’s ability to represent complex relationships and unique data modes depends on the number of parameters; if this number is too

low, information will inevitably be lost, and the resulting generated data may not represent the true distribution.

- **Functional Approximation Error:** Assuming infinite data (therefore no statistical approximation error), and a model with infinite capacity (therefore no functional expressivity error), we can still suffer errors from the training procedure itself. For example, optimisation may fail to converge to the optimal parameters of the distribution due to local minima or barren plateaus.

Each of these error types can individually stimulate the progression of model collapse; however, more often than not, they compound on one another and can be difficult to untangle. Moreover, other error types exist, such as floating-point approximation, which are assumed to be negligible in comparison (Shumailov et al. 2023).

Many papers (e.g., (Shumailov et al. 2023, 2024; Alemohammad et al. 2024)) introduce and demonstrate collapse dynamics for simple toy case Gaussian distributions. The focus in Shumailov et al. (2023), however, is to establish model collapse for Large Language Models and image data. They show how recursive fine-tuning of a pretrained transformer causes degradation across a selection of metrics for text data (wikitext2), and show a case example using a Variational Autoencoder on the MNIST dataset. Alemohammad et al. (2024) introduce their definition of model collapse (MADness), before introducing examples with Generative Adversarial Networks, Denoising Diffusion Probabilistic Models, as well as Normalizing Flows, all for image data: MNIST and FFHQ. Martínez et al. (2024) also shows collapse dynamics on image data (MNIST, Oxford 102 Flower, CUB-200-2011) for many of the same models. These examples are representative of the current literature in that they focus on image or text models. To the best of our knowledge, this is the first instance of a formal study of recursive self-training collapse for realistic tabular data specifically.

Synthetic Data Generation

In recent years, machine learning models and data mining tasks have become ever more complex, and while the first wave of data scarcity has seemingly been solved by self-supervision, better models, and more computational resources, a new roadblock has manifested itself in the form of data protection regulations such as the GDPR (European Parliament and Council of the European Union 2016). Microdata on citizens, customers, and patients is incredibly valuable for knowledge discovery tasks, but also very sensitive, why data protection is crucial in protecting the rights of individuals.

Synthetic data has been proposed as a possible mediator, helping to democratise the information locked away in sensitive data without violating privacy (Shanley et al. 2024). Synthetic data is artificially generated data that captures the structure, patterns, and characteristics of real data while ensuring the anonymity and confidentiality of individuals. In theory, the use of synthetic data is not limited to data privacy, but may also prove useful in combating other contemporary issues in data science, such as data scarcity, data fair-

ness, and data bias. As such, synthetic data may be applied to tackle data inequality by supplementing underrepresented groups in training datasets, thereby supporting the development of fair ML systems (van Breugel et al. 2021).

While the promise of synthetic data is immense, its efficacy largely depends on the quality and reliability of the generation process. Several widely recognised generative modelling frameworks exist, such as Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) and Variational Autoencoders (VAE) (Kingma and Welling 2014). Both have been adapted to tabular data with various augmentations (see e.g., (Xu et al. 2019)). Today, there exist dozens of augmentations of several generative frameworks specifically for tabular data (Hernandez et al. 2022; Murtaza et al. 2023).

Once generated, synthetic data is typically evaluated on various dimensions of *fidelity*, *utility*, and *privacy* (if applicable) (Lautrup et al. 2024). Fidelity captures modes of resemblance and quality, that is, how well the generated data matches the real data in terms of complex relationships of variables and other statistical properties. Utility measures how well the generated data can be used as a proxy in downstream tasks and yields a comparable and valid analysis. Privacy metrics measure if personally identifiable information is left accessible in the synthetic data; there is a certain overlap with overfitting metrics. We will not explicitly treat privacy in this work.

Statistical fidelity can be further subdivided into marginal measures (e.g., Wasserstein distance, Hellinger Distance, and the Kolmogorov–Smirnov statistic), measures of pairwise relationships (correlation- and/or mutual information differences), and measures of population resemblance (propensity MSE, Maximum mean discrepancy, etc.). For details, see the literature reviews (Dankar, Ibrahim, and Ismail 2022; Hernandez et al. 2022; Lautrup et al. 2024).

Besides the statistical fidelity measures, perhaps the most practically meaningful measure of utility is whether the synthetic data can provide a valid statistical analysis in the downstream task(s) for which it was created. Common methodologies include training a classification algorithm on synthetic data, and then validating its performance on held-out real data, which was not used to train the generative model, using accuracy measures, relative accuracy, AUROC, etc. (Hittmeir, Ekelhart, and Mayer 2019; Lautrup et al. 2024)). This paradigm is often referred to as Train-on-Synthetic-Test-on-Real (TSTR) (Esteban, Hyland, and Rätsch 2017).

Preliminaries

A prominent paper by Alaa et al. (2022) proposes a multifaceted fidelity measurement ‘ $\mathcal{E}$ ’ containing three metrics: α -precision, β -recall, and *Authenticity*, as important dimensions of synthetic data quality. First, α -precision measures resemblance, that is, whether the samples not only look realistic but also are typical. Second, β -recall is a measure of diversity, that the synthetic samples are able to capture the variability of real data. Lastly, *Authenticity* is a measure of the generated sample’s ‘originality’, meaning that they cannot be copies of the training samples (overfitting).

To accommodate the challenges associated with complex data types and multimodal distributions in the raw data space, the metrics are evaluated on embedded features learned from the real data. In the embedded space, both real and synthetic data are hyperspheres where typical samples are concentrated at the core, and higher variance samples are further away. These embeddings allow for a simple numerical approximation of the minimum volume subsets, and the resulting overlap can be understood intuitively in terms of precision-recall analogues (cf. (Sajjadi et al. 2018)).

Since we use elements from Alaa et al. (2022) in our demonstration below, we include a brief overview of the necessary concepts here, but refer the reader to the original paper for further details and justification.

First, let us define $\mathcal{S}^\alpha$, the α -support of a distribution $\mathbb{P}$ as the minimum volume subset of the support $\mathcal{S} = \text{supp}(\mathbb{P})$ that contains a probability mass of α :

$$\mathcal{S}^\alpha := \min_{s \subseteq \mathcal{S}} V(s), \quad \text{s.t. } \Pr(s) = \alpha, \quad (1)$$

where $V(s)$ is the volume of the set s and $\alpha \in [0, 1]$. We can define α -precision (P_α) as the probability that a synthetic sample is within the α -support of the real distribution $\mathcal{S}_r^\alpha$,

$$P_\alpha := \Pr(\tilde{X}_g \in \mathcal{S}_r^\alpha), \quad \alpha \in [0, 1], \quad (2)$$

where $\tilde{X}_g = \Phi(X_g)$ denotes the embedded representation of the synthetic data under the embedding function Φ . In the same way, we can define β -recall (R_β) as the probability that a real sample is within the β -support of the synthetic distribution $\mathcal{S}_g^\beta$,

$$R_\beta := \Pr(\tilde{X}_r \in \mathcal{S}_g^\beta), \quad \beta \in [0, 1]. \quad (3)$$

To calculate the metrics in practice, Φ is chosen so that it embeds the real data into a hypersphere with an isotropic mass distribution around the centre. The α -support can then be understood as a concentric shell that contains the α -support set, and we can calculate α -precision and β -recall for individual samples as binary conditions of whether they are within or without. That is,

$$f_P(\tilde{X}_g) = \begin{cases} 1, & \tilde{X}_g \in \mathcal{S}_r^\alpha, \\ 0, & \tilde{X}_g \notin \mathcal{S}_r^\alpha, \end{cases} \quad (4)$$

$$f_R(\tilde{X}_r) = \begin{cases} 1, & \tilde{X}_r \in \mathcal{S}_g^\beta, \\ 0, & \tilde{X}_r \notin \mathcal{S}_g^\beta. \end{cases} \quad (5)$$

Conducting a sweep over hyperspheres of different radii (i.e., all values of $\alpha, \beta \in [0, 1]$) allows us to define the mean absolute deviation of P_α and R_β as the integrals:

$$\Delta P_\alpha := \int_0^1 |P_\alpha - \alpha| d\alpha, \quad (6)$$

$$\Delta R_\beta := \int_0^1 |R_\beta - \beta| d\beta. \quad (7)$$

Clearly $\Delta P_\alpha, \Delta R_\beta \in [0, 1/2]$, which quantify the deviation between the P_α , and R_β curves and their optimal values.

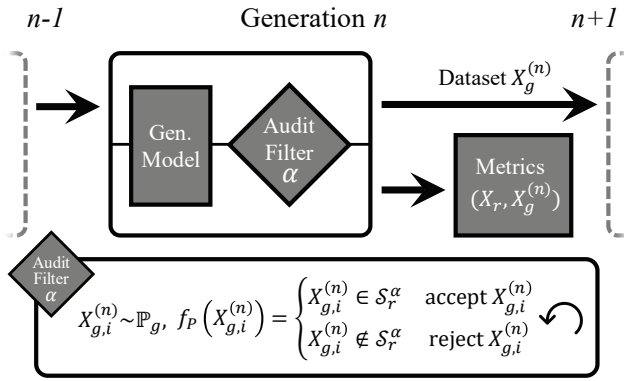


Figure 1: Generation Iteration Sketch. The figure shows the pipeline setup for the n th iteration of the experiment. The dataset from generation $n - 1$ is used to train the generative model. Next, the auditing filter iteratively selects outputs from the generative model (subject to α -condition eq. 4) until the required number is reached. The resulting dataset $X_g^{(n)}$ is evaluated against the original training data X_r (the input to generation 0) and forwarded to the next generation ($n + 1$). It is also shown how the auditing filter draws samples from the generated distribution and redraws if they do not meet the auditing criteria.

These numbers are usually transformed into the (integrated) P_α , and R_β scores, IP_α and IP_β :

$$IP_\alpha = 1 - 2\Delta P_\alpha, \quad IP_\beta = 1 - 2\Delta R_\beta, \quad (8)$$

which are scalar scores that range from 0 (worst) to 1 (best). It is these scores that we refer to as the general α -precision and β -recall measures going forwards.

Experimental Setup

For our study of model collapse by compounding errors across multiple generations of tabular generative models trained on their own synthetic outputs, we present the following experimental methodology.

The premise of the experiment is straightforward: The initial instance of the generative model (generation 0) is trained on the real data of the selected dataset. After training the model, it is used to generate a synthetic dataset $X_g^{(0)}$ of equal size to the initial training set, which serves as training data for the next model instance (generation 1). Importantly, each generation instantiates a new model, so each generation represents a full retraining rather than a fine-tuning of its predecessor. After the first generation, all subsequent generations ($n > 0$) are trained purely on the synthetic output $X_g^{(n-1)}$ of the previous generation (see Figure 1). Additionally, as is also seen on the figure, the experiments implement a post-hoc audit filtering which is used to reject low-quality synthetic samples; in the case of the “no audit” experiments, this element is not used¹.

In the auditing case, we adopt the auditing pipeline from Alaa et al. (2022), which is said to create “a curated synthetic dataset with an improved overall performance”. They

¹The code, raw results, and additional plots are available above.

Dataset	# Instances	# Features	# Classes
diabetes ²	768	9	2
wine ³	178	14	3

Table 1: Datasets Overview. The table summarises the datasets used in this study. The number of records and variables is shown. Please note that these are the values *after* preprocessing was applied.

show specifically how auditing can give a small but reliable boost to the performance on a downstream classifier trained on audited synthetic data. Alemohammad et al. (2024) implement a similar sampling bias for recursive training loops. They specifically test the dynamics of recursive retraining with these biases, evaluating only the basic precision and recall metrics (Sajjadi et al. 2018) across generations of their models working with image data.

The auditing pipeline we use acts as a filter on the generative model. The model generates samples, and the filter allows only those within the desired α -precision to propagate forward (sample-level condition eq. 4). In our experiments, we use three audit levels, $\alpha \in \{0.4, 0.6, 0.8\}$, in addition to the “no audit” base case, which allows us to compare the effect of the auditing over the course of our experiments.

For evaluation, after each generation, we gather measurements of the general α -precision and β -recall metrics (eq. (8)), the balanced accuracy of a TSTR experiment with a logistic regression classification model (we held out some of the real data for this evaluation), and we log the class retention rate (number of classes beyond one that is present in the synthetic data).

For our experiments, we choose two deep generative models, *CTGAN* and *TVAE* (Xu et al. 2019), accessed through the Synthcity suite (Qian, Cebere, and van der Schaar 2023). Across the two datasets shown in Table 1, we run 10 generations of our experimental pipeline for each of the three auditing levels and for the “no audit” baseline. To assess stochastic variability in the pipeline trajectories, each experiment is repeated across ten random seeds. Results are aggregated across seeds and reported as the mean with the unscaled unit of variation (standard deviation).

For 10 seeds, 2 models, 2 datasets, and 4 settings, we have 160 distinct simulations, and across 10 generations, this results in a total of 1,600 unique model instances trained. The experiments were performed on a machine with an NVIDIA GeForce RTX 3080 GPU, AMD Ryzen 7 5800X 8-Core CPU, and 16 GB of RAM.

Results and Analysis

In this section, we show the most important results from our experiments. For clarity, some of the plots shown feature only one dataset-model combination; however, similar figures for the other datasets and model combinations are available in the supplementary material¹.

²(Sigillito 1990)

³(Aeberhard and Forina 1992)

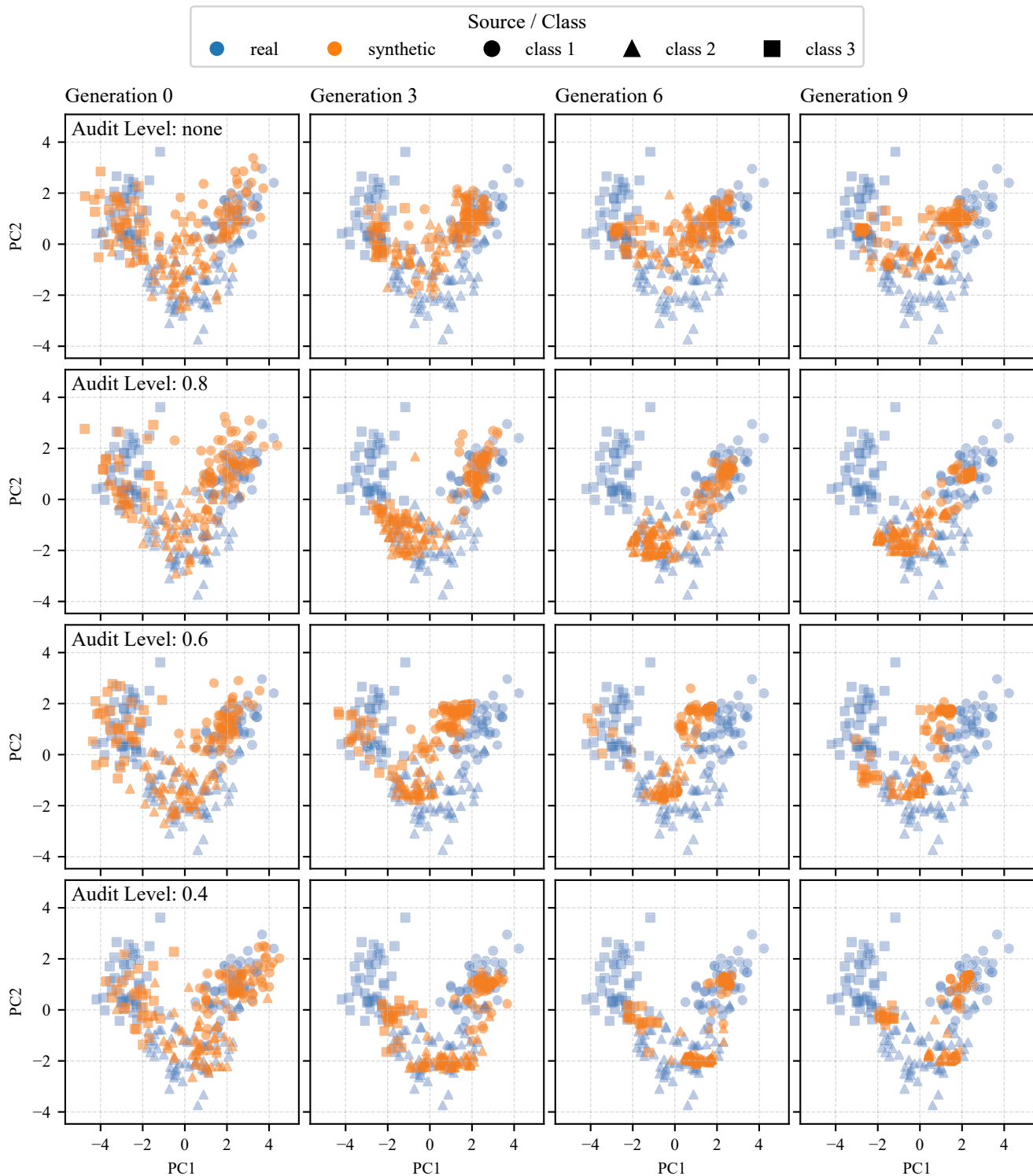


Figure 2: Principal Components and Synthetic Samples of the Wine Data (CTGAN). The grid of figures shows synthetic datasets sampled at different generations (on the horizontal) at different auditing conditions (shown vertically). The figure serves to give an intuition of what happens at the sample level. Most notably, groupings of synthetic data become more focused (diversity loss), and in one of the auditing cases ($\alpha = 0.8$), the “class 3” cluster is entirely lost (mode dropping). These projections are the result of a single simulation run; more quantitative evidence of the same observations is provided and discussed below.

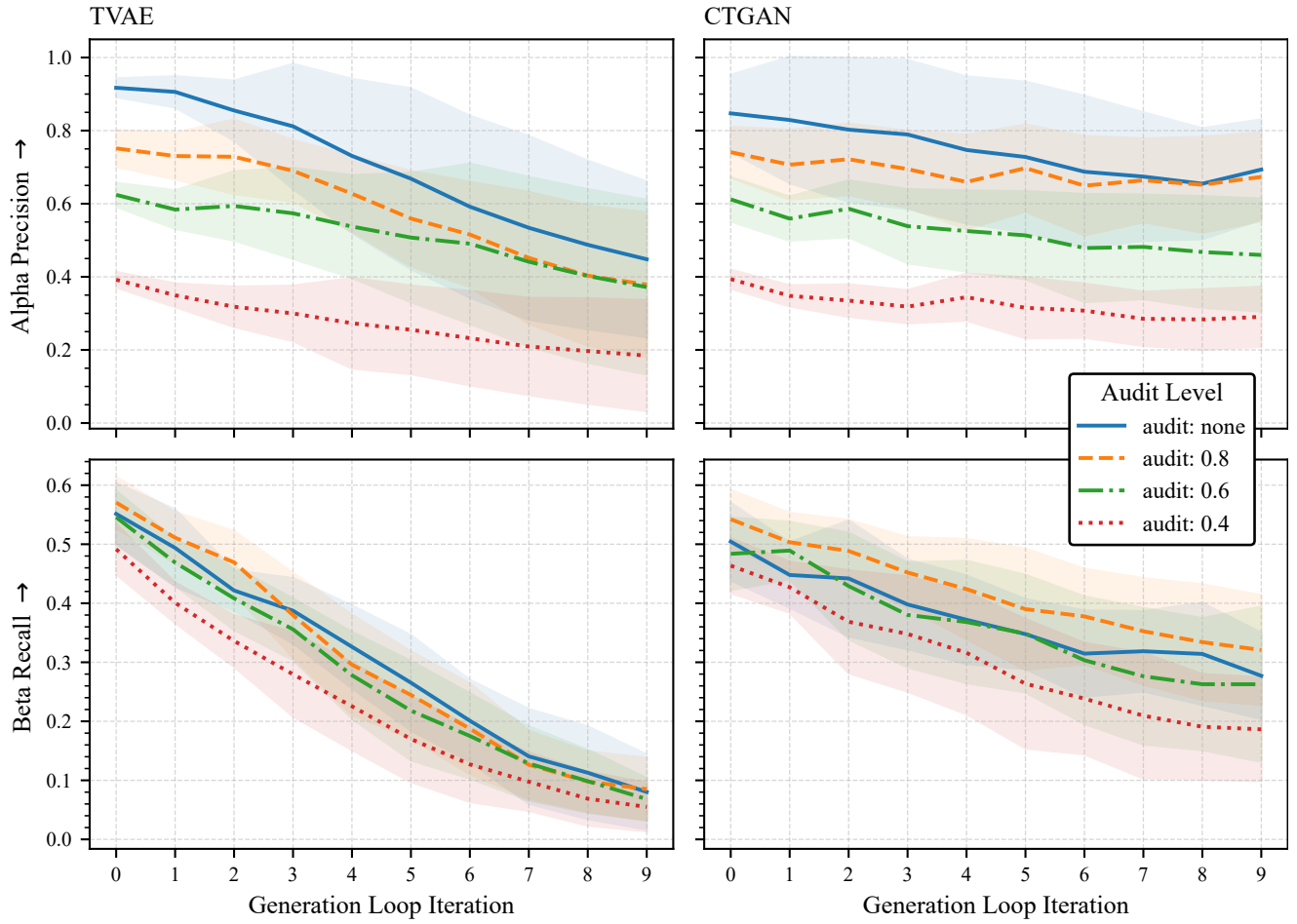


Figure 3: Distributional Metrics, α -Precision and β -Recall for Synthetic Wine Data. The figure shows the fidelity scores of synthetic data across generations for both TVAE and CTGAN models under each audit condition. The figure shows how both metrics deteriorate and that the initial scores approximately correspond to the audit levels. Please note that the y-axis measures refer to the general α -Precision and β -Recall (eq. 8) and *not* the sample-level conditions.

Before showing quantitative and aggregated results, we provide a first view of the dynamics at play in Figure 2. This figure provides a concrete visualisation of how the distribution of synthetic samples develops over generations under each of the four audit conditions. However, it should be interpreted cautiously. The projection is only a partial view of the data, based on the first two principal components fitted on the real data. It also shows the dynamics for a single seed, and therefore does not capture the stochastic variation across repeated runs.

Even with these limitations, the figure gives a valuable first indication of the failure mode for this model. At generation zero, each model is trained on purely real data, so the first column of Figure 2 shows model output before any recursive synthetic training has occurred. In each case, the initial synthetic data is visually consistent with the real distribution. However, as generations progress, the synthetic samples become increasingly concentrated in a smaller region of the principal component space. That is, the model appears to generate samples that remain locally plausible while pro-

gressively losing diversity.

This distinction is important for interpreting how the audit conditions may accelerate deterioration. The α -precision audit filters out samples that, for a given α , fall outside the estimated α -support of the real distribution. Such a filter is intended to remove low-fidelity samples. However, depending on the threshold, it may also remove tail observations or minority modes. Figure 2 suggests that this trade-off between sample-level precision and distributional diversity becomes more pronounced over repeated generations.

The class markers in Figure 2 show that some classes become increasingly under-represented and, in later generations, may disappear entirely. Even among the surviving classes, the variance collapse implies that some generated samples occupy regions that appear implausible under the ground truth class distribution. This suggests that recursive synthesis affects both the marginal feature distribution and the class-conditional structure of the data.

These observed patterns motivate the quantitative analysis that follows. The PCA plot suggests a loss of diver-

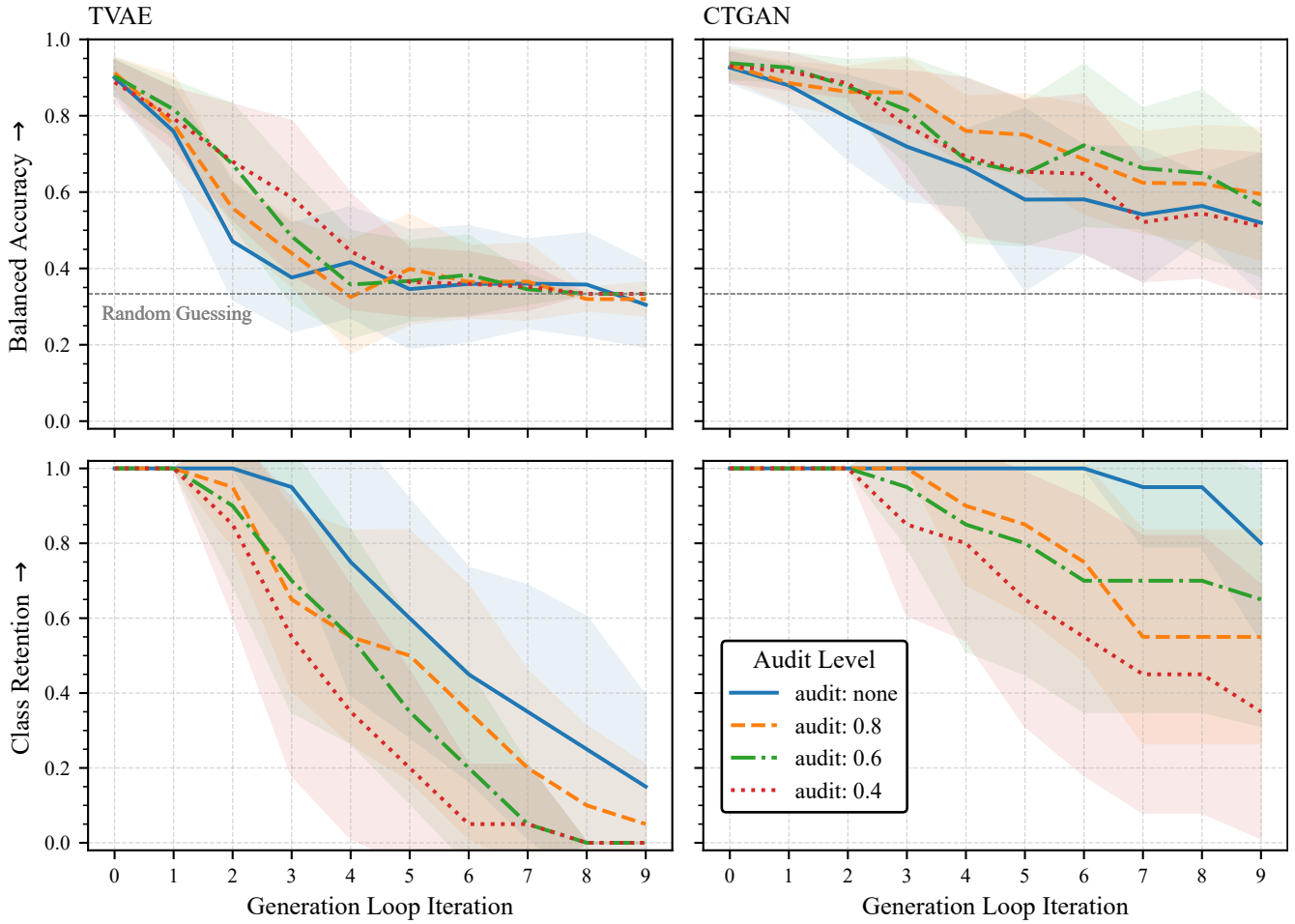


Figure 4: Prediction Metrics, Balanced Accuracy and Class Retention for Synthetic Wine Data. The figure shows the balanced accuracy of a TSTR classification task over the course of the generation loop iterations. In addition, we show the average class retention (fraction of classes beyond the first) across the repeated simulations. It is worth explicitly noting that the difference in class retention between the two models is quite significant.

sity, possible class collapse, and distributional shift for class-conditional structure, but it cannot fully establish these effects on its own. The following results, therefore, expand on this visual examination to quantitative changes in statistical fidelity and downstream predictive utility.

We repeat the experiment for 10 seeds and log evaluation results at every generation. The fidelity results presented in Figure 3 largely agree with the qualitative patterns observed in Figure 2 and provide evidence that they correspond to measurable changes in the distributional metrics. Across both models, recursive synthetic training leads to a decline in both the general α -precision and β -recall (eq. 8). The rate of decline varies between the models, with TVAE apparently degrading faster and more severely than CTGAN.

The negative slope of the α -precision corresponds exactly to the variance collapse we saw in Figure 3. Namely, as the synthetic samples are concentrated into a smaller and smaller region of the sample space, their density, relative to the real points, becomes greater; this disparity causes the in-

tegral over the hyperspheres of alpha support to trigger unevenly in eq. (6). Similarly, for the β -recall, which also declines, consistently with the variance collapse: The β -recall measures directly how well the synthetic distribution covers the real distribution, when the synthetic samples are highly localized, only the higher β -levels in eq. (7) (if any at all) support real samples causing a high discrepancy in the general β -recall value.

On the model level, the difference in behaviours of TVAE and CTGAN is particularly noteworthy, suggesting that the severity of the collapse effect depends on the generative model, but also shows that both models will eventually succumb to it. Remarkably, the TVAE model seems to suffer variance collapse irrespective of the auditing level, whereas for CTGAN, the auditing plays a larger role. Overall, Figure 3 shows that aggressive sample-level filtering can be harmful when synthetic data is repeatedly reused for training; more auditing increases the risk of curating away informant coverage, rather than just noise. The results look

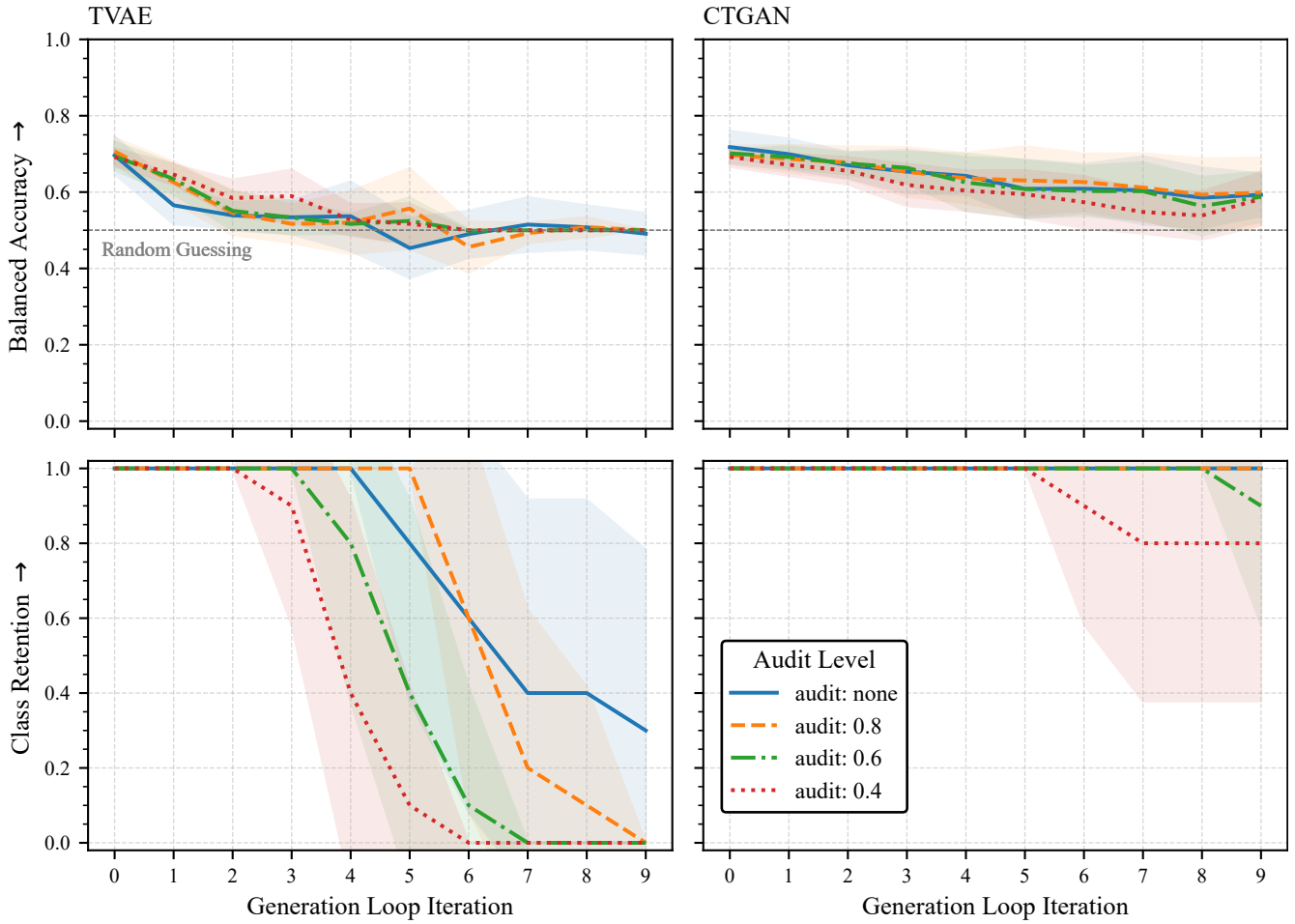


Figure 5: Prediction Metrics, Balanced Accuracy and Class Retention for Synthetic Diabetes Data. The figure shows the balanced accuracy of a TSTR classification task over the course of the generation loop iterations. In addition, we show the average class retention (fraction of classes beyond the first) across the repeated simulations. It is worth explicitly noting that the difference in class retention between the two models is quite significant.

similar for the diabetes dataset (shown in the supplementary material), with the notable exception that TVAE deteriorates more significantly on α -precision.

The balanced accuracy of the TSTR classification task and the class retention rate is shown in Figure 4 and Figure 5, for the wine and the diabetes datasets, respectively. These results display a similar degradation across generations for both models. The drop in accuracy is largely explained by the disappearance of all classes beyond one, as is seen in the class retention aggregate measure. Again, the comparison between the two models is perhaps the most obvious signal beyond the general deterioration. TVAE is again the weakest of the two, dropping all but one class and resorting to random guessing in fewer than ten generations on most seeds. CTGAN, on the other hand, manages to keep up for longer, and for the diabetes data, it retains the two classes on all but the most extreme auditing levels for the number of iterations performed here. For TVAE on the diabetes data, the class dropping seems strongly aligned with the degree

of auditing, where the stronger sample filtering also drops the second class the fastest on average. Another curious observation is that the order of the auditing level curves on the balanced accuracy almost looks opposite to the order in which they drop classes; however, the small differences inside the balanced accuracy curves are completely covered by the overlapping variances, so this pattern is unlikely to be generally reliable.

In the experiments, we keep the synthetic dataset at a fixed size equal to the original training data. Since the auditing filter may reject low-quality samples, this constraint necessitates a repeated procedure where the generative model is queried until enough samples meet the inclusion criteria (as described in Alaa et al. (2022)). In Figure 6, we show the proportion of samples that were rejected out of the total number of samples generated. The behaviour reveals that auditing takes its strongest effect in the first generations, and additionally, that stricter auditing leads to higher initial rejection. This result agrees with the intuition. While the real

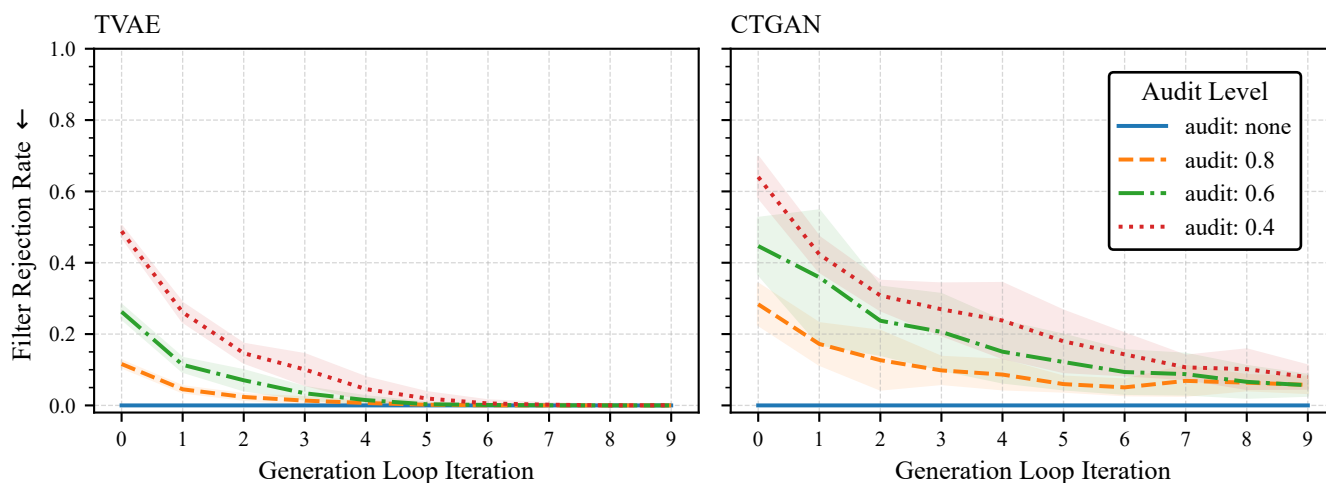


Figure 6: Auditing Filter Rejection Rates for Synthetic Diabetes Data. This figure shows how many samples are discarded in the auditing filter (Figure 1) process, expressed as the fraction of the number of samples rejected vs. the number of samples drawn before stopping.

data distribution that is used to audit against remains static, the variance of synthetic samples decreases, and they become concentrated in a smaller region of the sample space, meaning that the number of synthetic samples that fall outside of the α -support of the real distribution also shrinks. Hence, the long-run effects of auditing are that the model eventually internalizes the auditing filter and ceases to produce samples outside of that range (crucially, without being pushed to behave in this way that is!).

Discussion

With the current general lack of moderation of synthetic content, model collapse remains a highly relevant concern for the future development of machine learning algorithms and AI systems. Until now, major works have focused on language models and images, where the effect is readily observable, but limited work has investigated model collapse for tabular data. In this work, we contributed preliminary evidence for model collapse dynamics in tabular data. Our results demonstrate that recursive synthetic-data reuse can lead to diversity loss, mode dropping, and declining downstream utility across generations. Moreover, our experiments with sample-level auditing show that excessive post hoc curation may provoke the earlier onset of these afflictions. In the following, we initiate a discussion informing future best practices and guidelines on the matter.

Model Collapse in Tabular Data. At first glance, tabular data may appear less at risk of model collapse than other modalities. First and foremost, because the unrestricted sharing, sourcing, and aggregating of tabular datasets on the open internet is more niche than, say, text, images, and other media formats. Despite this, in certain domains, the associated risks of model collapse for this format may be higher (Wyllie, Shumailov, and Papernot 2024; Koul, Duran, and Hernandez-Boussard 2025). This is because of the pivotal role of tabular data across domains such as healthcare,

education, and public administration, where it supports decision systems that directly affect lives and livelihoods (Davernport and Kalakota 2019). The patterns observed in our experiments and in the related literature support that tails of the distributions are lost, meaning that any ensuing algorithmic unfairness is most likely to affect rare, minority cases. Healthcare is an example of a domain where the generation of accurate synthetic data would be especially valuable, as privacy often constrains the free use of real patient data. Also in healthcare, however, many useful learned decision systems rely on the accurate detection of minority cases or extreme instances. Here, synthetic data may be applied to enhance data equality in, for example, a biomedical dataset (e.g., (Frid-Adar et al. 2018)). However, as a cautionary example, Bhanot et al. (2021) reveals how even in prominently used and published first-generation synthetic datasets (MIMIC-III, and others), strong gender, age, and race biases remain, and especially vulnerable groups are already under-represented when measured against the source data. Future learning systems may pick up on such biases, and this risk could be further compounded if insights herefrom or if samples are directly recycled into later generations of data generators. This is particularly concerning because the failure mode can be difficult to detect. Even considering our experiments, the deterioration was subtle in the beginning, and unlike with synthetic images or text, where artifacts may be visible to the human eye, shifts in tabular feature distributions, class-conditional structure, or minority subgroup coverage can be easily and inadvertently overlooked.

The role of auditing pipelines further complicates these issues. The literature has suggested post hoc auditing as a way to filter low-quality generated samples out of generated results in order to improve the performance of downstream models trained on synthetic data (Alaa et al. 2022). In a single generation setting, such a result has been demonstrated to be true. In a recursive setting, however, any downstream utility gain must be understood to come at the ex-

pense of the diversity and precision of the synthetic distribution. Overly restrictive auditing is likely to remove too much variation from the training distribution, thereby harming all modes of data quality. Our prediction metric results (averaged across multiple runs), did not show a convincing increase in utility for any of the auditing conditions, datasets, or model choices, not even for the first few generations. Perhaps more dependable results could have been achieved with more moderate auditing levels; however, if this is the case, then we need to contemplate whether short-term gains in performance are justifiable against the future collapse risks.

Limitations. Some noteworthy limitations of this study include the use of only two datasets. Both are relatively small, and while they are typical examples of realistic datasets of this size, this number of evaluations and the sizes do limit the extent to which the observed dynamics should be expected to generalize across all tabular data. Secondly, the two models used, while being widely employed, are not representative of all synthesis methods for tabular data. Just as we observed CTGAN to be generally more robust to collapse dynamics when compared to TVAE, other models should similarly be expected to be more or less vulnerable to this phenomenon. Third, the experiments we conduct here implement a fully synthetic retraining loop; in contrast, we may also imagine other experimental methodologies, including branching paths, loops where synthetic data from all previous generations is used, and loops where fresh real data is included along with synthetic data (see e.g., (Alemohammad et al. 2024; Gerstgrasser et al. 2024)). Some of these methods could mitigate or exacerbate collapse phenomena; however, even with more parameters and dimensions to control, neither of these setups could perfectly model the full complexity of modern data ecosystems. Finally, we do not explicitly evaluate privacy or fairness metrics, even though these are central in motivating synthetic data in sensitive domains. We could speculate that empirical privacy measures would be improved during model collapse, save for the localized area of the sample space where the synthetic samples are concentrated. For similar reasons, we omitted to use the authenticity measure from Alaa et al. (2022), since it also would be showing a good signal for the wrong reasons.

Future work could explore how the model collapse dynamics extend to larger high-dimensional tabular datasets. In addition, more models should be explored, particularly other deep learning architectures (diffusion models, transformers, language models, etc.) and classical generative models, to see if some families of models are more dependable for maintaining synthetic data quality.

Recommendations. Despite these limitations, the results point to a broader concern. Much of the model collapse literature has focused on images and language, where deterioration is often easier to observe. Tabular data has received significantly less attention despite its ubiquity in institutional decision systems. Our results show that tabular generative models are not exempt from the many collapse phenomena observed for image and language models across the literature. With the slow creep of the collapse effects observed in our experiments, the outlook is indeed more bleak in the

short term, where the collapse effects may go entirely unnoticed. Compared to the more visually striking collapse effects exhibited with other modalities, collapse artifacts in tabular synthetic data are more difficult to detect. Future benchmarks should take care to conduct a multifaceted evaluation to monitor all dimensions of data quality, with a special emphasis on keeping diversity measures satisfactory and maintaining the integrity of dataset sub-groups (e.g., clusters and classes).

While ethical and epistemic issues associated with synthetic data have received increasing attention (see, e.g., (Koul, Duran, and Hernandez-Boussard 2025; Shanley et al. 2024)), frameworks aimed at safeguarding against model collapse remain understudied. Ethical and legislative efforts are needed to advise and direct future research, especially in domains where synthetic tabular data may influence institutional decisions. Additionally, if data collected from digital environments is to remain a reliable resource for future learning algorithms and AI systems, initiatives for marking and responsibly maintaining both *what is* but also *what isn't* synthetic data are essential efforts. Moreover, despite the adversarial feedback loop encountered by automatic detection systems and watermarking technologies for the moderation and control of artificially generated content, continued commitment to these technologies is not in vain. Finally, since different generative models behave differently when trained on previously generated synthetic outputs, research is also needed to determine if and why some models are better suited for mitigating model collapse.

Conclusion. To summarise, our results suggest that model collapse generalises to realistic tabular data settings. Practices such as model auditing (sample filtering) can cause earlier model collapse, which highlights a tension between oversight and long-term model reliability. Our findings underscore the need for careful governance; current legislative and ethical efforts should maintain the focus on transparency, data provenance, and moderation of artificially generated content to limit the risks associated with recursive training.

Acknowledgments

A.D.L. and L.G.P. acknowledge support from the MIST project, funded by the Novo Nordisk Foundation under grant reference number NNF25OC0103204.

References

- Aeberhard, S.; and Forina, M. 1992. Wine. UCI Machine Learning Repository.
- Alaa, A. M.; van Breugel, B.; Saveliev, E. S.; and van der Schaar, M. 2022. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *ICML 2022*, 290–306. Baltimore, Maryland, USA: PMLR.
- Alemohammad, S.; Casco-Rodriguez, J.; Luzi, L.; Humayun, A. I.; Babaei, H.; LeJeune, D.; Siahkoohi, A.; and Baraniuk, R. G. 2024. Self-Consuming Generative Models

- Go MAD. In *Proceedings of the 12th International Conference on Learning Representations*, ICLR 2024. Vienna, Austria: OpenReview.net.
- Bhanot, K.; Qi, M.; Erickson, J. S.; Guyon, I.; and Bennett, K. P. 2021. The Problem of Fairness in Synthetic Healthcare Data. *Entropy*, 23(9): 1165.
- Bieringer, S.; Butter, A.; Diefenbacher, S.; Eren, E.; Gaede, F.; Hundhausen, D.; Kasieczka, G.; Nachman, B.; Plehn, T.; and Trabs, M. 2022. Calomplification — the power of generative calorimeter models. *J. Instrum.*, 17(09): P09028.
- Chandra, B.; Dunietz, J.; Roberts, K.; Lee, Y.; Fontana, P.; and Awad, G. 2024. Reducing Risks Posed by Synthetic Content An Overview of Technical Approaches to Digital Content Transparency. Technical report, National Institute of Standards and Technology, Gaithersburg, MD.
- Chen, J.; Mowlaei, M. E.; and Shi, X. 2020. Population-scale Genomic Data Augmentation Based on Conditional Generative Adversarial Networks. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB'20, 26:1–26:6. Virtual Event, USA: ACM.
- Cloud, A.; Le, M.; Chua, J.; Betley, J.; Szyber-Betley, A.; Mindermann, S.; Hilton, J.; Marks, S.; and Evans, O. 2026. Language models transmit behavioural traits through hidden signals in data. *Nature*, 652: 615–621.
- Dankar, F. K.; Ibrahim, M. K.; and Ismail, L. 2022. A Multi-Dimensional Evaluation of Synthetic Data Generators. *IEEE Access*, 10: 11147–11158.
- Davenport, T.; and Kalakota, R. 2019. The potential for artificial intelligence in healthcare. *Future Healthc. J.*, 6(2): 94–98.
- Esteban, C.; Hyland, S. L.; and Rätsch, G. 2017. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. arXiv:1706.02633.
- European Parliament; and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).
- Frid-Adar, M.; Klang, E.; Amitai, M.; Goldberger, J.; and Greenspan, H. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. In *Proceedings of the 15th IEEE International Symposium on Biomedical Imaging*, ISBI'18, 289–293. Washington, DC, USA: IEEE.
- Gerstgrasser, M.; Schaeffer, R.; Dey, A.; Rafailov, R.; Sleight, H.; Hughes, J.; Korbak, T.; Agrawal, R.; Pai, D.; Gromov, A.; Roberts, D. A.; Yang, D.; Donoho, D. L.; and Koyejo, S. 2024. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. arXiv:2404.01413.
- Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. arXiv:1406.2661.
- Hernandez, M.; Epelde, G.; Alberdi, A.; Cilla, R.; and Rankin, D. 2022. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493: 28–45.
- Hittmeir, M.; Ekelhart, A.; and Mayer, R. 2019. On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, ARES 2019, 29:1–29:6. Canterbury, UK: ACM.
- Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, ICLR 2014. Banff, AB, Canada: OpenReview.net.
- Koul, A.; Duran, D.; and Hernandez-Boussard, T. 2025. Synthetic data, synthetic trust: navigating data challenges in the digital revolution. *Lancet Digit. Health*, 7(11): 100924.
- Lautrup, A. D.; Hyrup, T.; Zimek, A.; and Schneider-Kamp, P. 2024. Systematic Review of Generative Modelling Tools and Utility Metrics for Fully Synthetic Tabular Data. *ACM Comput. Surv.*, 57(4): 90:1–90:38.
- Lundberg, E.; and Mozelius, P. 2025. The potential effects of deepfakes on news media and entertainment. *AI Soc.*, 40(4): 2159–2170.
- Martínez, G.; Watson, L.; Reviriego, P.; Hernández, J. A.; Juárez, M.; and Sarkar, R. 2024. Towards Understanding the Interplay of Generative Artificial Intelligence and the Internet. In *Epi UAI*, volume 14523 of *Lect. Notes Comput. Sci.*, 59–73. Springer.
- 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. *Comput. Sci. Rev.*, 48: 100546.
- Park, S.; and Nan, X. 2026. Generative AI and misinformation: a scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation. *AI Soc.*, 41(2): 1501–1515.
- Qian, Z.; Cebere, B.; and van der Schaar, M. 2023. Synthcity: facilitating innovative use cases of synthetic data in different data modalities. arXiv:2301.07573.
- Sajjadi, M. S. M.; Bachem, O.; Lucic, M.; Bousquet, O.; and Gelly, S. 2018. Assessing Generative Models via Precision and Recall. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, NeurIPS 2018, 5234–5243. Montréal, Canada.
- Schaeffer, R.; Kazdan, J.; Arulandu, A. C.; and Koyejo, S. 2025. Position: Model Collapse Does Not Mean What You Think. arXiv:2503.03150.
- Shanley, D.; Hogenboom, J.; Lysen, F.; Wee, L.; Lobo Gomes, A.; Dekker, A.; and Meacham, D. 2024. Getting real about synthetic data ethics : Are AI ethics principles a good starting point for synthetic data ethics? *EMBO Rep.*, 25(5): 2152–2155.
- Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Gal, Y.; Papernot, N.; and Anderson, R. J. 2023. The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv:2305.17493.

Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; and Gal, Y. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759.

Sigillito, V. 1990. Pima Indians Diabetes Database. UCI Machine Learning Repository.

van Breugel, B.; Kyono, T.; Berrevoets, J.; and van der Schaar, M. 2021. DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, NeurIPS 2021, 22221–22233. Virtual Event, USA.

Wyllie, S. C.; Shumailov, I.; and Papernot, N. 2024. Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024*, 2113–2147. ACM.

Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019. Modeling Tabular data using Conditional GAN. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, NeurIPS 2019, 7333–7343. Vancouver, BC, Canada.