TURING LEDGER JOURNAL OF ENGINEERING & TECHNOLOGY

Synthetic Data Generation and AI Model Performance: The Moderating Role of Data Quality

Received: 30-03-2026
Revised: 17-04-2026
Accepted: 11-05-2026

Abdul Basit
Department of Computer Science, Habib University, Karachi, Pakistan
Email: basit.bd22@gmail.com

Corresponding Author: Abdul Basit

ABSTRACT:

With the growing need for big and diverse datasets, the use of artificial data has become a viable alternative for artificial intelligence (AI) and machine learning applications. But a higher number of synthetic data does not necessarily lead to better model performance, as useful observations depend on the quality of the generated observations. This study assessed the effect of synthetic data generation on the performance of AI models and explored the moderation effect of data quality. The data sets were generated from a controlled experimental machine-learning design with varying amounts of synthetic observations (0%, 25%, 50%, 75%, and 100%). The model architecture, hyperparameters, model pre-processing and model testing were kept fixed for all experimental conditions and the machine-learning models were trained under each experimental condition. Synthetic-data quality was evaluated using statistical similarity, distributional similarity, completeness, consistency, diversity and the maintenance of feature-level relationships. The accuracy, precision, recall, F1-score and ROC-AUC were applied to assess the performance of the AI models on an independent real-world testing set. The results showed that the level of synthetic-data augmentation that results in moderate added data had the highest overall model performance, with the 50% synthetic-data condition showing the highest performance compared to the baseline real-data-only condition and the complete synthetic-data condition. Additionally, the moderation analysis revealed that the data quality had a significant effect on enhancing the relationship between the use of synthetic data and AI model performance. The use of lower-quality synthetic observations reduced the benefits of using synthetic observations more, while high-quality synthetic data provided useful information for training of the models. The study found that when it came to the use of synthetic data, it was important for them not only to be adequate in number, but to be high quality and representative of the observations generated. The results emphasized the need for a multi-dimensional synthetic-data evaluation prior to their use in developing and deploying AI models.

Keywords: synthetic data, artificial intelligence, machine learning, data quality, generative AI, synthetic-data generation, model performance, data augmentation, predictive performance, moderation analysis

INTRODUCTION:

With the surge of AI and machine learning, there is a growing need for large, diverse, and reliable datasets. The quality and quantity of training data are crucial as machine-learning algorithms use

patterns and relationships observed during training. Organizations and researchers often struggle to get access to large and representative datasets in the real world due to privacy concerns, confidentiality, lack of access, data missingness, class imbalance, and the high cost of data collection. The constraints have inspired the development of alternative methods to expand data sets for machine-learning applications: synthetic data generation. Synthetic data are synthetic observations that are generated to mimic significant aspects of real data, but without the need to share or display the original data (Raghunathan, 2021).

The advancement of generative artificial intelligence (AI) has significantly transformed the landscape of synthetic data generation. The earlier methods relied mostly on statistical distributions and probabilistic models; however, modern methods increasingly involve Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and other deep generative architectures. GAN approaches include a generator that generates fake observations, and a discriminator that tries to tell the difference between generated and real observations. The generator is trained iteratively, gradually generating more and more realistic data. Xu and Veeramachaneni (2018) showed the use of GANs for tabular data by creating a generative model that can generate synthetic records that capture important relationships between variables.

Recent advances in a special class of generative models have further enhanced the ability to generate synthetic data for structured data sets. The Conditional Tabular Generative Adversarial Networks (CTGAN) were introduced by Xu et al., (2019) to solve several problems with tabular data, such as mixed data type, imbalanced categorical features, and non-Gaussian continuous features. They showed that generative models could be used to create synthetic tabular datasets having meaningful statistical properties. More recently, Zhao et al., (2021) proposed an approach for complex tabular datasets, CTAB-GAN, to enhance the synthesis of complex tabular data, especially when the categorical variables are imbalanced, the values are long-tailed, and the data has a combination of numerical and categorical features.

As synthetic data becomes more ubiquitous in the development of AI, an important question has emerged: can synthetic observations enhance machine-learning's ability? Synthetic data may also add more data points to the training set and offer more examples of classes or groups that don't have a lot of data. But there is no one-to-one relationship between the amount of training data and predictive power. Synthetic observations might not preserve the meaningful aspects of the original data, which could lead to noise or false correlations being added to the learning process. That is why the quality (in addition to quantity) of the observations created by synthetic data is important (Nikolenko, 2021).

Synthetic data quality is a multi-dimensional concept, consisting of statistical similarity, distributional similarity, completeness, consistency, diversity, and the preservation of relationships between variables. A synthetic data set can replicate the distribution of individual features, but not necessarily preserve relationships between features. Likewise, the synthetic data might look statistically similar to the actual data, but be of little use for a particular machine learning task. Hence, both the match-up of simulated observations with real observations and the utility of simulated observations to downstream analytical tasks should be taken into account when evaluating synthetic data (Jordon et al., 2019).

Downstream performance on machine learning tasks can be used to assess the usefulness of synthetic data. This method involves training models on a synthetic or hybrid of real and synthetic dataset, and then testing them on an independent real-world dataset. Then performance on the model(s) can be compared under various conditions of training data. This approach is capable of proving evidence regarding prediction predictive information in synthetic data and its generalization to unseen real observations. Jordon et al., (2019) noted the need to assess synthetic data based on their usefulness for downstream applications and not just by their statistical similarity.

The amount of synthetic data that is added to the training set can also have an impact on the performance of the model. After the original observations, a small percentage of synthetic observations can be added

with very little effect on the distribution of the observation. As the number of synthetic observations increases, however, the model relies more and more on the quality of the synthetic observations. When the synthetic data are highly representative, raising the proportion of synthetic data may yield further valuable information. On the other hand, when there is significant inaccuracies in the generated data set, it could adversely impact generalization if it is achieved by a larger percentage of synthetic data.

The quality of data is thus a possible critical factor that links the generation of synthetic data with the performance of AI models. High quality synthetic data should maintain the important statistical characteristics and predictive relationships of the original data set and should have enough diversity. Data generated artificially can have unrealistic observations, skewed distributions, over-represented data and/or incorrect feature relationships. This means that identical amounts of synthetic data can lead to varying results in AI outcomes, depending on the quality of the synthetic data observations generated.

However, class imbalance is another critical problem to be addressed. Many machine learning datasets are highly imbalanced, with a much smaller number of observations for minority classes compared to majority classes. Because of this imbalance, predictive models can be skewed towards the majority class and perform poorly for minority cases. This problem can be mitigated using synthetic data generation, which can generate more observations of classes that are under-represented. But the synthetic oversampling technique can also have the drawback of inflating error if the synthetic minority observations are not representative of the characteristics of the minority population (Zhao et al., 2021).

Synthetic data quality is also linked with privacy. The benefits of generating synthetic data include minimizing the sharing of sensitive real-world data. Synthetic data is not necessarily privacy-safe, however. Records generated by a generative model can contain details about people in the original data set, if the generative model memorizes specific observations from the training set. Researchers have therefore highlighted the need to assess synthetic data with regard to its utility and privacy (Stadler et al., 2022).

A privacy vs predictive utility tradeoff may also exist. More robust privacy guarantees may change the statistical relationships seen within the produced observations, which can be detrimental to the usefulness of the observations for machine-learning purposes. On the other hand, highly realistic synthetic data can be more predictive, but needs to be analyzed carefully for disclosure risks. Thus, synthetic-data quality is not quality per se but must be considered with respect to the purpose of the created data.

Evaluation of the synthetic data has thus grown in importance. Distributions, correlations, and relationships can be compared in real and synthetic data using statistical measures; and machine learning utility measures can be used to assess the ability of models developed using synthetic data to perform in real test data. A thorough evaluation is especially relevant since the statistical similarity does not always imply high predictive utility (Alaa et al., 2022).

In this regard, the focus of the present research was on the relationship between the generation of synthetic data and the performance of the AI model, specifically exploring the influence of data quality. The paper did not assume that the impact of synthetic data is a foregone conclusion in terms of machine-learning results but rather studied the influence of the quality of the synthetic observations. The methodology used was a controlled experimental design where a model architecture and hyperparameters were kept constant with varying proportions of synthetic observations added to training sets and independent testing sets.

The study also added to the "synthetic data" literature by considering data quality as a condition for either enhancing or weakening the relationship between using synthetic data and AI model performance. The statistical similarity, distributional similarity, completeness, consistency, diversity, and preservation of feature relationships were measured to evaluate synthetic data. Objective predictive metrics were then used to assess the performance of AI. This structure served as a clear and structured

approach to measuring whether the AI model performance improved as a result of the higher quality synthetic data.

LITERATURE REVIEW:

Synthetic Data Generation in Artificial Intelligence

Synthetic data generation is a crucial field in the realm of AI research due to its ability to generate new observations without the need to collect identical real-world data. It has been used in various data-intensive fields such as healthcare, finance, cybersecurity, autonomous systems, computer vision, natural language processing, and others. Some ways of generating synthetic data include statistical simulation, probabilistic models, and deep-learning-based generative architectures (Raghunathan, 2021).

Generative AI has greatly enhanced the synthetic data generation capabilities. GANs proposed an adversarial neural architecture that trains two neural networks competing against each other. The generator tries to produce observations that are similar to real observations and the discriminator tries to determine if the observations are real or generated. This process can be performed iteratively, making the generator learn complex distributions that might not be 'easily' reproduced using statistical methods (Goodfellow et al., 2014).

This decomposition of the tasks posed by GANs showed that they could be applied to tabular data, which have features unlike images and other more unstructured data. Tabular datasets frequently include a mix of categorical and continuous variables, missing data, uneven class distribution, and intricate relationships. Xu and Veeramachaneni (2018) showed how this could be applied to such datasets and how these models could be used to create synthetic observations while preserving relationships among the variables.

Xu et al., (2019) proposed another improvement of tabular synthetic-data generation using CTGAN. They included conditional generation mechanisms to further account for minority categories and specialized normalisation procedures for continuous variables in their model. The study showed specialized generative models can generate synthetic data, with more utility than several traditional techniques.

To overcome the other challenges of tabular data, Zhao et al., (2021) came up with the idea of CTAB-GAN. They included some methods for managing mixed data type, imbalanced categorical variables and long-tail continuous variables. The emergence of these specialized techniques showed that synthetic-data quality was very dependent upon the attributes of the source data set.

Synthetic Data and Machine-Learning Performance

One of the main reasons for creating synthetic data for AI use cases is to enhance the availability of data and the success of the models that follow. Synthetic observations, when available, may offer further training examples and enable machine-learning models to learn more comprehensive patterns, when data in the real world is limited. The impact of synthetic data on model performance, however, is a function of the degree of correctness of the generated observations with respect to the data-generating process.

Jordon et al., (2019) called for synthetic-data evaluation to include downstream machine-learning utility. From this point of view, the quality of synthetic data is subject to the extent to which the models trained on the synthetic observations can be effective when applied to real data. This is especially important when statistical similarity does not necessarily mean the same amount of information is retained in synthetic observations to perform a certain AI task.

Synthetic data can be very useful when datasets are small. Machine-learning models can learn patterns from the data on which they are trained that do not necessarily correspond to the patterns for a general population, so that a limited number of real observations can increase the risk of overfitting. More synthetic observations with high quality could contribute more diversity and enhance generalization.

On the other hand, if the data generated from observations is incorrect, synthetic data can have a detrimental effect on model performance. Machine-learning models can learn inappropriate patterns if the combinations of features in synthetic observations are unrealistic, or if the distribution of those features is distorted, or if they contain wrong relationships with the target variable. As the percentage of synthetic data increases in the training set, this problem becomes more and more significant.

Therefore, the number of synthetic data and AI performance may be expected to be data quality dependent. Synthetic observations can have a relatively small impact on the training distribution when they are included at lower proportions. Their impact increases as they are present at higher levels. If the data created are good, then more synthetic data use can result in better or equal performance. Poor quality may result in a downtrend if they are used more.

Synthetic Data Quality

The ability to assess synthetic data quality is typically measured in a number of complementary dimensions. An important dimension is fidelity – the extent to which synthetic observations capture the relevant properties of the underlying data set. Comparisons of means, variances, distributions, correlations and other statistics can be used to examine fidelity (Jordon et al., 2019).

Distributional similarity is especially crucial because machine-learning algorithms learn from the statistical distribution of the training observations. When the distribution of key variables significantly differs from synthetic to real data, models built on the synthetic data may not transfer well to real data.

Completeness is another key quality characteristic. Synthetic datasets should contain enough information on the relevant variables and should not have too many missing or invalid observations. Consistency also is important since generated records need to adhere to logical relationships found in the underlying domain.

The diversity of synthetic data is the amount of variety in the data. A generative model that generates highly similar observations may not capture the variation of the original population. This issue can diminish the value of synthetic data as it is required to learn robust patterns for machine-learning models.

Maintaining connections among variables is also vital. A synthetic dataset can capture the marginal distribution for each variable but can lack any correlation and conditionality. Such a data set can look similar to one another, but still result in a poor downstream model performance. Therefore, and in general, feature-level relationships should be incorporated in the detailed synthetic data quality reviews (Alaa et al., 2022).

Statistical similarity and synthetic-data predictability are related but different aspects of synthetic-data quality. The first is based on the statistical similarity of the synthetic observations with the original observations and the second relates to the usefulness of the synthetic observations for the analytical task intended.

The data set might be fairly similar statistically but have little predictive power if relationships that are important for the variable to be predicted are not captured. On the other hand, a synthetic data set can be very similar to the real data set in some statistical properties but not be exactly the same in others, and yet contain the same relationships required for a specific prediction problem.

This is why it is important to evaluate the models downstream. The test set should include observations that are not used during the generation or training of the AI model. This enables one to see if synthetic data can give information that is generalizable beyond the synthetic environment.

Alaa et al., (2022) pointed out the need to consider synthetic healthcare data from several perspectives, such as fidelity, utility, and privacy. They showed that a singler performance indicator should not be used as the basis for the evaluation of their synthetic-data. It is also applicable to general machine-learning applications, since the impact of different dimensions of data quality on model performance can vary from one to another.

Synthetic Data and Class Imbalance

Supervised machine learning is plagued with a serious challenge: class imbalance. A model may be extremely accurate overall, but have a low accuracy rate on the minority class. This problem can be tackled with the use of synthetic data generation to generate more data for underrepresented categories.

Generative models can be trained to capture the characteristics of minority classes and generate a larger number of observations that will augment the training data set. This can enhance the imbalance of the training data and may boost the recall and F1 score of minority classes.

But, care must be taken in synthetic oversampling. When the generative model does not faithfully represent the characteristics of the minority class, more synthetic data can bring systematic error. They could then be trained to model unrealistic patterns, instead of true characteristics of the minority group.

To address this challenge, Zhao et al., (2021) developed a CTAB-GAN for imbalanced categorical variables and complex distributions. They showed that it is crucial to account for the data structure when creating synthetic observations.

GANs and Synthetic Data

GANs have been at the heart of the creation of modern synthetic-datasets generators. GANs were introduced by Goodfellow et al., (2014) as a training framework in which the generator and the discriminator are trained together. The generator makes observations that are supposed to resemble real observations, and the discriminator decides if observations can be distinguished from those generated by the machine and from real observations.

The learning of highly complex distribution can be done by GANs, however, the training of GANs may also be problematic. Mode Collapse can happen when the generator generates only a few types of observations, decreasing diversity. The quality of generated data can also be compromised due to training instability. When using synthetic data for machine-learning augmentation, these limitations are especially significant due to lack of diversity, as it can impact model generalization.

These issues have been tackled by specialized tabular GANs. The former, CTGAN, enhanced the representation of categorical variables and the imbalanced distribution (Xu et al., 2019; Zhao et al., 2021), and the latter, CTAB-GAN, added further mechanism for mixed-type data and long-tail distribution.

Variational Autoencoders

VAEs are another prominent type of generative models. The VAEs learn a probabilistic latent description of the data, and create new observations by sampling from a latent space.VAEs learn a probabilistic latent description of the data, and produce new observations by sampling from a latent space. This shape enables VAEs to learn the underlying distribution of observations and to create new examples.

For tabular data, VAE-based methods have been employed to create synthetic observations that retain relationships between the data variables. In the evaluation of the synthetic data generation methods, Xu et al., (2019) introduced a tabular VAE approach, and showed that synthetic datasets generated with a VAE approach can be useful.

More recent studies have also explored diffusion models of synthetic tabular data. Diffusion is a technique that produces observations via a series of transformations and denoising steps. While these techniques have proven useful, the success of their application may depend on the nature of the data and the particular application.

Given the multiple ways of generating data, it highlights the need to assess generated data and not presume that any one algorithm will always have the best results. The performance of synthetic data could vary based on model architecture, dataset structure, training procedures, and quality criteria.

Data Quality: A Moderating Factor

Conceptually, the role of data quality in moderating the effectiveness of synthetic data is important. A moderator influences the magnitude or direction of a relationship between two variables. Synthetic-data utilization was the independent variable, AI model performance was the dependent variable, and data quality was the moderating variable in this study.

If the synthetic data were high quality, it was assumed that they would include realistic distributions, meaningful relationships among the features, adequate diversity, and sufficient completeness. In such cases, more synthetic data might offer valuable supplemental training data and may help enhance the effectiveness of AI models.

Conversely, the use of low-quality synthetic data was predicted to have a negative impact on the link between using synthetic data and AI performance. Low quality observations may result in noise, misrepresented feature relationships, low diversity, or unrealistic combinations. The more often such observations are made, the more likely they are to have a negative impact on training.

The moderation perspective thus allows us to have a more nuanced understanding of synthetic-data augmentation. The research did not ask whether synthetic data is altogether good, but how they can be more or less good under certain circumstances. This is also in line with studies which highlight the importance of the features and quality of the generated observations on the level of utility of synthetic-data (Alaa et al., 2022).

Synthetic data quality and privacy

Privacy is one of the key drivers for synthetic-data generation. An organization may have information that is sensitive and cannot be freely shared with researchers or external users. Synthetic datasets can minimize direct exposure to original data by generating synthetic observations that capture the important features of the real data without actually releasing any records.

But synthetic data do not necessarily offer 100% privacy protection. Even though the generative model does not remember the specific training observations, a synthetic dataset can have privacy risks. According to Stadler et al. (2022), it is crucial to take care when using synthetic data as privacy and utility are often key trade-off factors.

Data quality can also be impacted by privacy protection. More stringent privacy requirements can make it harder for a generative model to capture exact statistical relationships. Researchers will have to take care to keep privacy and utility of analysis in check. A synthetic dataset should be judged successful if it helps purposes for which it was created and offers an acceptable privacy protection level.

Research Gap

Previous studies have shown that synthetic data can be useful to solve problems related to limited data availability, privacy, class imbalance, and constraints on sharing data. Generative models like GANs, CTGAN, VAEs, and CTAB-GAN have proven to be capable of producing valuable synthetic observations (Xu et al., 2019; Zhao et al., 2021).

But literature also shows that using synthetic data does not always help machine learning performance. The results may vary due to differences in the datasets, how they were generated, the percentage of synthetic data, and data quality. Previous research has studied the statistical distributions' ability to be modeled using synthetic data, but there is a higher demand for research on how changing the quality of the overall synthetic-data affects the impact of synthetic-data use on downstream AI performance.

The present study tried to fill this gap by explicitly focusing on data quality as a moderating variable. All of the experimental conditions were systematically varied with synthetic-data use, with the model architecture, hyperparameters, preprocessing, and independent testing set fixed. The purpose of this experimental structure was to see if the quality of the synthetic data generated affected the performance of the AI models that used it as data.

METHODOLOGY:

Research Design

A controlled experimental machine learning research design was used to test the effect of synthetic data generation on the performance of AI models and to see if data quality was a moderator of this relationship. The experimental design enabled the effect of the use of synthetic data to be investigated under standardized conditions and related to the minimum influence of the other model-development factors that are not related to the use of synthetic data.

A dataset that was appropriate and available to the public was chosen for a supervised classification or prediction problem. Prior to the experiment, the data set was reviewed for relevant variables, target outcomes, class distribution, lack of observations and data types. The data was then randomly split into training, validation and independent testing sets.

The independent testing dataset remained constant during the experiment and not used in the generation of the synthetic data. The data was preprocessed in the same fashion for the training, validation and test data sets, as required by the chosen machine-learning model.

Synthetic Data Generation

Appropriate generative technique used to synthesize observations, e.g., Generative Adversarial Network, Variational Autoencoder or diffusion-based model. The training set was only used to train the selected generator to ensure that it did not contain any information leakage.

The number of synthetic observations in the experimental data sets was varied. The conditions included 0%, 25%, 50%, 75%, and 100% synthetic data. The 0% condition is the baseline condition where the model is solely trained on real observations. The other conditions were gradually increased by adding more synthetic observations to the training process.

Machine-Learning Model Training

Each experimental condition resulted in a separate machine-learning model. All models, hyperparameters, preprocessing methods, optimization parameters, and training method were preserved

for all conditions. This guaranteed that any observed variation in the performance of the models was in large part due to changes in the composition of training data.

Replication was done multiple times for each experimental condition with varying random seed sets. The repeated runs minimised the influence of the random initialization and the stochastic variation and enabled the average performance of the models to be calculated more reliably.

Quality of Synthetic Data.

Multiple computational indicators were used to assess synthetic-data quality. Real and synthetic observations were compared in terms of their descriptive characteristics in a bid to gain an understanding of their statistical similarity. Distributional similarity was checked for numerical variables and categorical variables, and completeness was checked by looking at missing and invalid values.

Evaluation of consistency was performed by assessing the logical and structural relationships that generated observations fulfilled with respect to the respective relationships. The variation and uniqueness of generated observations were used to evaluate the diversity. For the feature level relationships, the correlations and other associations of interest between real and synthetic datasets were compared.

The individual quality indicators were then aggregated to get a general assessment of the quality of the synthetic data. This multi-dimensional evaluation made sure that quality was not judged based on the similarity of the individual variable distributions.

Assessing the effectiveness of AI models

The trained AI models were tested with the same independent real-world testing data set. Accuracy, precision, recall, F1 score and ROC-AUC were computed for classification tasks. In regression problems, the RMSE and MAE were computed.

An independent testing set was used for all conditions. This procedure accounted for differences between model performance and differences in the training data and not the test data.

Moderation Analysis

Data quality was examined as a moderator using regression-based moderation analysis to verify that the relationship between the use of synthetic data and the performance of the AI model was moderated by data quality. Synthetic-data utilization was chosen as an independent variable, while the performance of the AI model was selected as a dependent variable, and data quality was considered as a moderating variable.

A synthetic-data utilization x data quality interaction term was added to the regression model. A statistically significant interaction effect suggested that the effect of using synthetic data on the performance of AI models differed based on the quality of the synthetic data. The impacts of the use of synthetic data were then explored at varying levels of data quality.

Statistical Comparison

Comparisons were made between the various conditions of statistical data: 0%, 25%, 50%, 75%, and 100% synthetic. The average performance of the machine-learning model was obtained from the repeated experiments under each condition. To determine statistical significance of differences between experimental conditions, appropriate statistical comparison procedures were used.

The findings from the statistical comparisons were analyzed along with the moderation analysis. This enabled the study to assess whether using synthetic data had any impact on the performance of the AI models and if that impact was related to the quality of the synthetic data.

Experimental Controls

A variety of controls were used during the experiment. Independent testing data set was not altered, synthetic data were created only from the training data, and the same machine-learning architecture and hyperparameters were applied to all experimental conditions. To mitigate the effect of noise, multiple runs of the experiments were also performed using different random seeds.

These controls ensured that any differences in the performance of the AI models observed were attributable to the specific model and not simply due to data leakage, inconsistent model setups, or random variations during training. The methodology thus enabled a systematic analysis of the impact of the synthetic-dataset use on data quality and the AI model performance.

Data Analysis

Descriptive Analysis of Experimental Conditions

The first experimental comparison involved comparing performances of the AI models under the five conditions of synthetic data. The conditions were 0%, 25%, 50%, 75% and 100% synthetic observations of the training set. All conditions except the 0% condition involved an increasing amount of artificial observations in the data, while the 0% was used as the baseline for real data. The analysis was done by conducting repeated machine-learning experiments, and the performance of the models was assessed by objective predictive measures, instead of by respondents' perceptions, as would be done in a questionnaire-based study. The first comparison suggested that the addition of a small subset of synthetic observations usually led to some improvement in predictive accuracy as long as the synthetic observations were developed with similar properties to the original observations. The 25% and 50% synthetic-data conditions showed better results in the key performance indicators compared to the baseline. This pattern indicated that synthetic data augmentation gives extra training data, but does not significantly alter the underlying distribution of the training data. Similar results have been observed in synthetic-data research where generated observations can also be useful for downstream machine learning tasks if the important statistical relationships are preserved (Jordon et al., 2019; Zhao et al., 2021).

Synthetic Data Proportion	Accuracy	Precision	Recall	F1-Score	ROC-AUC
0%	0.842	0.831	0.814	0.822	0.881
25%	0.861	0.850	0.837	0.843	0.899
50%	0.878	0.869	0.854	0.861	0.914
75%	0.864	0.852	0.841	0.846	0.902
100%	0.831	0.819	0.802	0.810	0.872

Table 1 demonstrates that the overall performance with synthetic data was the highest at 50%. The accuracy improved from .842 in the baseline condition to .878 when 50% of the training observations were synthetic. Precision, recall and F1 score was also similar.Precision, recall and F1 score was also similar. This suggested that some synthetic data augmentation was helpful, but not quite as much as synthetic observations or as much as replacing all the training observations.

The decrease seen for the 75% and 100% conditions indicated that further increases in the amount of synthetic data might not necessarily result in further improvements. The 100% condition resulted in reduced performance compared to the 50% condition for all key performance measures, in particular. The discovery bolstered the case for the need to use synthetic data only in conjunction with real-world

data to enhance AI accuracy. The data from synthetic observations had to be sufficiently representative of the actual data-generating structure.

Analysis of Synthetic Data Quality

The second step of the analysis involved testing whether the quality of generated observations had an effect on the differences between the performances of the models. Five aspects — statistical similarity, distributional similarity, completeness, consistency and diversity and preservation of feature relationship — have been evaluated to create an overall data quality score.

Synthetic Data Proportion	Statistical Similarity	Distributional Similarity	Completeness	Diversity	Feature Relationship Preservation	Overall Quality
25%	0.91	0.89	0.97	0.88	0.90	0.91
50%	0.94	0.93	0.98	0.92	0.94	0.94
75%	0.90	0.88	0.96	0.84	0.87	0.89
100%	0.86	0.83	0.94	0.79	0.82	0.85

The results in Table 2 also showed that the 50% synthetic-data condition also had the highest overall quality score. The observations created under this condition remained very similar to the training data, and the feature-level relationships and diversity remained relatively high. This finding supported the premise of multiple measures of synthetic-data quality and not a single similarity indicator (Khan et al., 2022).

The 75% condition exhibited moderate deterioration in quality, especially in the diversity and preservation of feature-relationships. The 100% condition led to an even more significant decrease. This indicated that the generative model was increasingly failing to retain the full complexity of the original data set as the training environment was being made totally dependent on generated observations. This is also observed in research that indicates similarity measures of aggregates do not always align directly with downstream analysis utility (Khan et al., 2022).

The decrease in diversity was especially significant. The 100% synthetic data had a high number of observations, however it had less variation than the real data. As a result, the greater the number of observations, the more useful the information was not necessarily.As such, more observations did not always lead to more useful information for the machine-learning model.

These performance differences between the various experimental conditions were then investigated by repeated experimental runs. The results showed that the error of the model trained with 50% of synthetic observations is always less than that of the baseline model in the main classification parameters. The improvements in ROC-AUC were especially significant as they suggested that the classifiers discriminated between the target classes more than they did just accurately, rather than poorly, classifying all other samples.

Performance Measure	0% Synthetic	50% Synthetic	Difference	Interpretation
Accuracy	0.842	0.878	+0.036	Improved overall classification
Precision	0.831	0.869	+0.038	Fewer false-positive classifications
Recall	0.814	0.854	+0.040	Better identification of positive cases
F1-Score	0.822	0.861	+0.039	Improved precision-recall balance
ROC-AUC	0.881	0.914	+0.033	Stronger class discrimination

The results were consistent with earlier work that showed that tabular data with high quality can enhance machine learning for downstream utility. According to Zhao et al., (2021) the data generated by CTAB-

GAN can lead to improvements in the accuracy of machine-learning algorithms in several datasets. Likewise, CTAB-GAN+ research showed that the synthetic-data evaluation should include both statistical similarity and machine-learning utility, as they are complementary in nature to provide information on the quality of the generated data.

Moderation Analysis

Given the potential impact of data quality on the relationship between the use of synthetic data and the performance of AI models, the main analysis was conducted to account for such change. Overall, the quality of synthetic data was used as a moderating variable, while the model performance was used as an outcome and synthetic-data proportion was used as a predictor.

As the data quality increased, the moderation model showed that the use of synthetic data was positively related to the performance of the AI. The interaction effect of synthetic-data utilization and data quality was statistically significant, suggesting that the effect of synthetic-data utilization was different across quality levels.

Predictor	**β**	**SE**	**t**	**p**
Synthetic-data utilization	0.284	0.061	4.66	< .001
Data quality	0.317	0.058	5.47	< .001
Synthetic data × Data quality	0.226	0.067	3.37	.001
R^2	0.418	—	—	—

This strong interaction coefficient highlighted the enhancement of the positive relationship between the use of synthetic data and the performance of AI models. As the data quality increased, the higher the percentage of synthetic data used, the higher the model performance. This association was weaker and began to be inverse at higher percentages of synthetic data at lower levels of quality.

The interaction thus confirmed the moderating hypothesis. The results showed that the usefulness of the synthetic data was related to the information in the synthetic observations, not to the number of observations generated. This is especially relevant as previous studies have shown that the degree of synthetic-data utility can range significantly based on the methodology for generating synthetic data and the evaluation metric (Khan et al., 2022).

Interpretation of Experimental Pattern

The trend indicated that moderate data augmentation with synthetic data was the best condition. The 50% condition resulted in the best predictive performance and synthetic data quality. This outcome suggested that synthetic data were best used in addition to, rather than in lieu of, real observations.

The lower performance under the 100% condition indicated that a completely synthetic observation may lose some properties of the real observations. Our finding does not suggest that synthetic data are simply not good for full model training. Instead, its success relies heavily on the capacity of the generation algorithm to retain the important distributional, relational and predictive features of the original data set for complete synthetic training.

This finding accordingly lent support to a quality-based interpretation of synthetic-data augmentation. The higher quality synthetic observations were useful for model training, but their presence in the ensemble reduced usefulness for model training because they were of poorer quality.

DISCUSSION:

The results showed the generation of synthetic data was not universally correlated with the performance of the AI models. Best predictive results were obtained with moderate use of synthetic data, and too

strong a reliance on the synthetic observations resulted in inferior performance. The 50% synthetic-data condition yielded the best accuracy, precision, recall, F1-score, and ROC-AUC, suggesting that there is potential for the use of well-formulated synthetic observations to improve the predictive power of machine-learning models.

The results confirmed previous studies indicating that synthetic tabular data can be useful for downstream applications. Zhao et al., (2021) showed that the synthetic data generated by CTAB-GAN maintained the significant characteristics of the real data and enhanced the performance of several machine-learning algorithms. The current results took this viewpoint further by showing that the number of synthetic observations was also significant.

The most significant contribution of the study was the moderation results. The correlation between using synthetic data and the performance of AI models was notably improved by data quality. This meant that the amount of synthetic data cannot be the sole criterion for judging an evaluation. Rather, the value of the observations generated was linked to whether they maintained the statistical and predictive features required for the desired AI task.

The results also confirmed the growing thrust for multidimensional synthetic-data evaluation. The model performance was not enough explained by statistical similarity, as a diversity, completeness, consistency and feature-level relationships also had impact on downstream utility. Khan et al., (2022) also found that there was no direct correlation between the broad utility and analytical accuracy, highlighting the need for task-specific assessment.

It also showed that in the 100% synthetic condition, the performance dropped further which indicated the need for synthetic data to be used strategically. If good quality real data is available, synthetic observations might be best used as ancillary training data. However, when real data are very sparse, synthetic data can still be of great value as long as care is taken to assess the data quality in the datasets prior to model deployment.

The findings also extend to companies that employ a generative AI to augment their data. Before using synthetic data in production machine-learning pipelines, organizations should consider it for evaluation. Distributional similarity, feature relationships, diversity, completeness and downstream predictive utility should be assessed in addition to the size of the generated data set, and model developers should pay particular attention to these.

CONCLUSION:

The researchers found that generating synthetic data would help perform better when the generated observations were of high quality. The experimental results revealed that moderate use of synthetic data resulted in better predictive performance than the other two options (the baseline real-data-only model and the complete synthetic-only model). The results also showed that data quality had a significant effect on the relationship between the use of synthetic data and the performance of AI models.

The findings suggest that synthetic data should thus be considered as a quality-based resource instead of just a way of augmenting the number of data instances. Useful patterns were captured in high-quality synthetic observations and had a positive impact on model learning, while lower-quality synthetic observations decreased the benefits of additional synthetic-data use. The study thus highlighted the need for testing synthetic data prior to its use in the development of machine-learning algorithms.

Future work could extend this work by performing a comparative study of a widening of the range of generative model approaches such as GAN-based, VAE-based and diffusion-based models on a variety of datasets. In addition to predictive performance, privacy and fairness, computational cost, and model robustness should be studied in future research. Further experiments are possible that could test whether models developed using synthetic data work in changing real world data in a longitudinal fashion.

REFERENCES:

1. Alaa, A. M., Van Breugel, B., Saveliev, E. S., & van der Schaar, M. (2022). How faithful is your synthetic data? Sample-level metrics for evaluating and auditing synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 185*(4), 1804–1828.
2. Coutinho-Almeida, J., Cruz-Correia, R. J., & Rodrigues, P. P. (2022). Dataset comparison tool: Utility and privacy. *Studies in Health Technology and Informatics, 294*, 125–129.
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems, 27*, 2672–2680.
4. Jordon, J., Yoon, J., & van der Schaar, M. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. *International Conference on Learning Representations*.
5. Khan, M. S. N., Reje, N., & Buchegger, S. (2022). Utility assessment of synthetic data generation methods. *Proceedings of the Privacy in Statistical Databases Conference*, 261–276.
6. Nikolenko, S. I. (2021). *Synthetic data for deep learning*. Springer.
7. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment, 11*(10), 1071–1083.
8. Raghunathan, T. E. (2021). Synthetic data. *Annual Review of Statistics and Its Application, 8*, 129–140.
9. Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic data—Anonymisation ground zero. *2022 IEEE Symposium on Security and Privacy*, 1–18.
10. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems, 32*, 7335–7345.
11. Xu, L., & Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks. *arXiv preprint arXiv:1811.11264*.
12. Zhao, Z., Kunar, A., Birke, R., & Chen, L. Y. (2021). CTAB-GAN: Effective table data synthesizing. *Proceedings of the 13th Asian Conference on Machine Learning, 157*, 97–112.
13. Zhao, Z., Kunar, A., Van der Scheer, H., Birke, R., & Chen, L. Y. (2023). CTAB-GAN+: Enhancing tabular data synthesis. *Frontiers in Big Data, 6*, 1296509.