

---

# DISTRIBUTED CONDITIONAL GAN (DISCGAN) AND SYNTHETIC HEALTHCARE DATA GENERATION

---

**David Fuentes**  
School of Data Science  
University of Virginia  
dmf4ns@virginia.edu

**Diana McSpadden**  
School of Data Science  
University of Virginia  
hdm5s@virginia.edu

May 4, 2022

## ABSTRACT

Modeling distributions of discrete and continuous tabular data is a non-trivial task with high utility. We propose discGAN: a generative adversarial network (GAN) to create synthetic tabular data specific to the healthcare domain using a distributed algorithm. Modeling non-Gaussian multi-modal data is addressed. We evaluated our GAN using machine learning efficacy, the Kolmogorov–Smirnov test (KS test) for continuous variables, and chi-squared test for discrete variables.

## 1 Introduction

### 1.1 Healthcare data: privacy concerns and utility

Healthcare information falls into a unique category of protected health information under the Health Insurance Portability and Accountability Act (HIPAA). This poses unique challenges to healthcare research, healthcare administration research, and open science. Anonymous, publicly-available healthcare information exist, such as MIT’s eICU Collaborative Research Database [1–3]. However, only 2,500 intensive care unit stays are represented in the publicly available, demo version of the database; thus, the ability to generate statistically similar synthetic data is useful for open-science research.

The complex dynamics between hospital stay characteristics – such as pre-existing conditions, medications, and patient demographics – may contribute to individual patient outcomes, and efficient and effective use of hospital resources. Using data science methods to understand these complex dynamics requires sufficient data to develop algorithms, and data science algorithms require adequate statistics across the parameter space. Tabular healthcare data includes continuous features, binary and multi-class discrete features, and could even include medical images. We address a tabular GAN for discrete and continuous variables. Generating statistically-similar hospital and patient data conditioned on hospital and medical characteristics has high utility for healthcare research.

### 1.2 Synthetic healthcare data: a solution?

The healthcare research community has been able to effectively close the door on public patient-scale research due to a lack of anonymous healthcare data. The contributions that could be made in hospital resource utilization, anomaly identification in patient care or hospital usage, and characterizing patient risk factors are significant, but only possible if enough data is available to feed data-hungry algorithms. We propose that GANs present a potential solution to these problems. With robust GAN architectures, researchers could leverage anonymous synthetic data for public health benefits without exposing the underlying individuals in the raw data.

## 2 Literature Survey

### 2.1 CTGAN: Modeling Conditional Tabular Data [4] from Synthetic Data Vault [5]

From MIT, CTGAN is the SOTA tabular data generator with the ability to generate discrete and continuous values, model non-Gaussian multi-modal data, and learn conditional distributions for many dimensions. [4] builds on previous tabular GAN implementations such as VeeGAN [6], TableGAN [7], and MedGAN [8]. CTGAN's contributions include "augmenting the training procedure with mode-specific normalization" to address multi-modal continuous columns. For each continuous column a variational Gaussian mixture model estimates the number of modes, then for each record in the column the probability of the value coming from each mode is calculated, then the sampled mode from the probability density is used for normalization of the value. Ultimately, the column value is represented as a scalar and a one-hot vector indicating which mode the value is from.

Additionally, CTGAN adds a conditional generator and training-by-sampling to address a lack of samples from minor categories during training, i.e. unbalanced data. CTGAN samples in a way "that all the categories from discrete attributes are sampled evenly (though not necessarily uniformly)" during training. CTGAN learns a conditional vector indicating which condition should be sampled. During training the generator is penalized for generating data that does not match an indicated conditional vector. CTGAN's training-by-sampling approach ensures that the model "evenly explore[s] all possible values in discrete columns", even minor categories. For benchmark datasets of real data, *machine learning efficacy* was used to evaluate performance. A set of test data was held out during training, and classification and regression models were trained on generated data, then tested on the holdout data. Classification tasks used accuracy and F1 metrics, and regression tasks used  $R^2$ .

### 2.2 "Md-gan: Multi-discriminator generative adversarial networks for distributed datasets" [9]

The earliest of the papers reviewed describing architectures for distributed GANs, the approach in [9] proposes a single, centralized generator, and multiple, distributed discriminator workers. Md-gan considered a scenario of  $N$  workers, possibly geographically distributed across datacenters, each with their own local "truth" training dataset (assumed i.i.d.). Local datasets remain on their worker, which limits the bandwidth needed for Md-gan when compared to an architecture that would require dataset sharing.

Md-gan evaluated its performance on communication and computation complexity, memory, and quality against a standalone GAN and FL-GAN – a distributed GAN with both a generator and discriminator on each worker that had their model weights averaged throughout training. Md-gan was solely evaluated on generating images. Md-gan was competitive on communication complexity for smaller batch sizes, less than 550, but with non-image data, batch size may behave differently. The paper reports an expectation that Md-gan has half the computation complexity of FL-GAN due to a single generator vs. a generator per worker; the complexity is not  $1/N$  because the consequence of the single, central generator is more frequent interactions between the workers and the generator.

[9] evaluated performance generating images from the MNIST, CIFAR10, and CelebA datasets using Inception Score (IS), and Fréchet Inception Distance (FID). [9] describes IS as an evaluation of the generated data when classified on a pre-trained Inception classifier. IS evaluates whether the generated data are "well recognized" and diverse. A higher IS score is better. FID measures a "distance between the distribution of the generated data and real data." To evaluate this distance, an Inception network or a network adapted from an Inception network is used, making this metric image specific.

Additionally, Md-gan evaluated GAN performance with and without peer-to-peer discriminator swapping throughout training. There was only a marginal gain in performance when periodically swapping discriminator weights between workers. Fault tolerance to worker crashes was also described. The authors report "the MD-GAN architecture manages to learn fast enough so that crashes, and then the removal of dataset shares, are not a problem performance wise."

### 2.3 "FeGAN: Scaling Distributed GANs" [10]

[10] demonstrates a distributed, conditional GAN with distribution of both the generator (G) and discriminator (D) on  $K$  devices. Different devices have access to different samples from different subsets of the total classes. Devices are asked to contribute based on their capabilities, i.e. memory and compute power. Devices must communicate to the server their capability and the number of samples per class. The central server will "load control" requests to devices to balance the contribution of device G's and D's trained on different class distributions with differing sample counts.

A central server starts each round of learning by selecting a group of devices from  $K$ . Devices are selected to "ensure a balanced number of samples per class". Only a subset of the total devices may be used during a training round. The server also provides selected devices with the current model state. The selected devices train for a configured number of

epochs, then send updated models to the server. After receiving updated G's and D's the server aggregates the models using a "KL weighting scheme" by "applying the softmax function on the KL divergence score of each device." This scheme promotes device model updates with data distributions similar to the global data distribution, and penalizes device model updates with greater divergence from the global data distribution. This method attempts to avoid mode collapse.

[10] evaluated performance generating images from the MNIST, Fashion-MNIST, and ImageNet using Fréchet Inception Distance (FID). [10] reports that the co-location of G and D on devices, KL weighting for model aggregation, and balanced sampling result in a generator powerful enough to generate multimodal distributions, demonstrated by their lower FID values compared to a vanilla federated learning GAN.

## 2.4 "Synthetic Learning: Learn From Distributed Asynchronized Discriminator GAN Without Sharing Medical Image Data" [11]

[11] demonstrated a distributed GAN with a central generator and distributed discriminators, similar to [9], but with a more complex architecture. Not only is each generated image identified as real or fake, but additionally [11] describes how "[t]he discriminator individually quantifies the fake or real value of different small patches in the image." The approach is termed Distributed Asynchronized Discriminator GAN (AsynDGAN). In [10], AsynDGAN's architecture was also used to generate medical images used as the sole training data for a segmentation model making this paper the first to validate GAN-generated data for its utility in medical research.

[11] proposed a GAN framework to mimic a multi-hospital system with each hospital providing a discriminator with access to its own unique true images of differing quantities. As [11] states, "data from different [discriminator worker] nodes are often dissimilar." The authors propose that AsynDGAN provides a method to unify the different datasets by the generator learning from multiple, likely disjoint, discriminators. An interesting finding of [11] is that an individual discriminator was only able to learn a single mode Gaussian due to its local information limitation, but with distributed discriminators, AsynDGAN was able to capture global information and multiple modes.

[11]'s evaluation of its generated images was focused on how well the generated images performed as training data for segmentation models of medical images, specifically, brain tumor and nuclei segmentation. AsynDGAN was compared on Dice score, Sensitivity, Specificity and 95% quantile of Hausdorff distance (HD95). Dice and HD95 provide two different measures of the distance between two subsets. A higher Dice score is better, while a lower HD95 score is better. Real brain tumor images and AsynDGAN-generated brain tumor images were evaluated in segmentation experiments. Compared to AsynDGAN-generated images, real brain tumor images had higher Dice scores (0.7485 vs 0.7043) and lower HD95 values (12.85 vs 14.94), but AsynDGAN-images had better scores than a non-distributed GAN, or any of the individual GANs trained from a single discriminator in the distributed architecture. This indicates that AsynDGAN improves upon a single discriminator architecture and that the distributed discriminators with disjoint real data provide valuable information to the distributed GAN.

## 3 Method

### 3.1 Identification of novel dataset

In order for GAN-generated synthetic data to be a viable open source of healthcare data, the underlying studies must be reproducible. In this vein, we reviewed publicly available datasets on which to train our GAN. For evaluation of the GAN synthetic data, the study must have a sufficiently large sample size so that the study can be reproduced on a subset of the total data in order for a validation set to be held out.

We propose leveraging existing knowledge of tabular GANs, conditional tabular GANs, and distributed GANs to produce a first step towards a generative model capable of producing statistically accurate, but variable tabular data from a small initial data set using distributed training. The demo eICU [2] database provides 2,500 patient-unit stay records and includes patient demographics such as age, ethnicity, and gender, past medical history, medications, initial diagnosis, treatments, and discharge diagnosis. A unit stay is the unique combination of a patient, the specific hospital ICU unit in which they stayed (e.g. Cardiac ICU, Neuro ICU, and Med-Surg ICU), and the start time for the patient in the ICU unit.

In developing discGAN and comparing to the state-of-the-art (SOTA) CTGAN, we evaluated the following features:

1. age: the age of the patient when they began their unit stay
2. hospital discharge offset: the number of minutes from unit admit time to when the patient was discharged from the hospital

3. CHF: indicator that the patient had a past history of congestive heart failure (CHF)
4. COPD\_nolimitations, COPD\_moderate, COPD\_severe: indicator that the patient had a past history of mild, moderate, or severe chronic obstructive pulmonary disease (COPD)
5. NoHealthProblems: indicator that the patient had no past history of health problems
6. asthma: indicator that the patient had a past history of asthma
7. homeoxygen: indicator that the patient had a past history of requiring oxygen treatment at home
8. hypertensionrequiringtreatment: indicator the patient had a past history hypertension (high blood pressure) requiring treatment
9. restrictivepulmonarydisease: indicator the patient had a past history of restrictive pulmonary disease
10. All [past history count]: count of all past history diagnosis entered into the eICU system when the patient was admitted to the hospital unit. These include past histories not listed above.
11. ethnicity: ethnicity of the patient. Ethnicity categories are: African American, Asian, Caucasian, Hispanic, Native American, and Other/Unknown
12. gender: gender of the patient. Gender categories are: Male, and Female.
13. discharge status: whether the patient was alive, or expired when they were released from the hospital unit. Discharge status categories are: Alive, Expired, and Other.

### 3.2 Model architecture review

We will use methods described in [4] to produce synthetic, single-table synthetic data. In particular, CTGAN overcomes several issues related to using GANs to generate synthetic tabular data; namely, "the need to simultaneously model discrete and continuous columns, the multi-modal non-Gaussian values within each continuous column, and the severe imbalance of categorical columns." [4] In contrast to "vanilla" GANs, CTGAN uses mode-specific normalization to overcome the issue of multi-modal, non-Gaussian distributions, and address data imbalance with a conditional generator and training-by-sampling. The overall architecture implements "fully-connected networks and several recent techniques to train a high-quality model." [4] Additionally, CTGAN addresses the issue of class-imbalanced data when training the generator, which is unusual in a typical GAN architecture. CTGAN aims to "resample efficiently in a way that all the categories from discrete attributes are sampled evenly [during training] [...]" and to recover the (not sampled) real data distribution during testing." [4] Thus, the data generation is conditional on each particular column value.

### 3.3 discGAN Approach

We generate tabular healthcare data using a distributed, conditional GAN, an appropriate tool to address our motivation because, in its simplest implementation, a GAN directs a zero-sum game in which one model, the generator, produces synthetic data and another, the discriminator, determines whether the generated data are true or erroneous. This process happens iteratively, with both models training based on the outcome of each game. In our selected domain, a GAN must be able to generate data for selected conditions; for example, it must handle not only the patient age distribution for the entire hospital, but the patient age distribution specific to the Cardiac ICU. This requires a GAN that can produce conditioned synthetic data. A conditional tabular GAN is trained similarly to a basic tabular GAN, with the additional necessity of learning the data distributions by class.

We call our implementation discGAN for *distributed conditional GAN*.

#### 3.3.1 Data Pre-Processing

discGAN expects a TensorFlow DataSet object. As such, our code required minimally pre-processed eICU data. We imported the eICU data via a CSV file into a pandas data frame, standardized the continuous variable, and one-hot encoded all discrete variables using `sklearn.preprocessing`, converting each feature to a NumPy array with each transformation. Lastly, we concatenated the scaled continuous arrays and one-hot encoded discrete arrays into a TensorFlow DataSet object.

#### 3.3.2 discGAN Architecture

discGAN is relatively lightweight. In particular, we forewent adding a residual block(s) to our model, which is opposed to the SOTA CTGAN implementation, as our model performed well without one. Both discriminator and generator utilize a standard Adam optimizer to update the model weights.

The discGAN generator contains an input layer followed by three sets of: (1) a dense layer, (2) leaky ReLU activation, and (3) batch normalization. After this series, a final leaky ReLU activation is incorporated before a final dense layer to create the output. Overall, there are 18,945 trainable parameters and 384 non-trainable parameters in the generator.

The discGAN discriminator has an even simpler architecture. Its input layer is followed by two triplets of (1) a dense layer, (2) leaky ReLU activation, and (3) dropout regularization before a final dense output layer. Overall, there are 7,105 trainable parameters and 0 non-trainable parameters in the discriminator.

Our model is malleable in its distributing capabilities. It is built so that it can be run without distributing, with either the generator or discriminator distributed, or with both distributed. We successfully provisioned multiple GPUs as multiple workers. In a true distributed environment, we imagined that dozens or hundreds of workers may be provisioned and we expect our methods to scale as described in [10], [9], and [11]. discGAN’s implementation of distribution and our distribution of the SOTA CTGAN is describe in section 4.2.

### 3.4 Evaluating generated tabular data

**Qualitative evaluation** Initial model development of 1D and 2D GANs utilized qualitative and simple quantitative evaluation of generated data. Intra-training distribution plots as seen in Figure 1 were used for evaluation. These plots demonstrate the GAN learning the data distribution(s); the generated data becomes visually similar to the true distribution throughout training. In addition to subjective comparison of histograms, the generated minimum, maximum, mean, and standard deviations of real and generated data were compared.

**Quantitative evaluation** To compare 2D and  $n$ -dimensional tabular data, we compared distribution of continuous and discrete columns from synthetic and real data. For continuous columns we use the two-sample Kolmogorov-Smirnov (KS) test [12]. The two samples are the real and the synthetic data. The KS test value is 1 minus the KS Test D statistic, which indicates the maximum distance between the real cumulative distribution function (CDF) and the synthetic CDF. [5] The average KS Test value for all continuous columns in the real vs. generated data is used as a statistical metric of the performance of the GAN on continuous columns.

For discrete columns, the chi-squared test compares the real and generated discrete distributions. The p-value of the chi-squared test is used to indicate the probability that the two discrete columns are from the same distribution. [5] The mean probability for discrete columns is returned as the CS Test metric.

Additionally, we compare the ability of the generated vs. real data to be used in machine learning algorithms. [5]’s *machine learning efficacy* [5, 13] metrics indicate if the generated data is comparably useful as training data for a machine learning task. We compare the performance of a decision tree classifier (Decision Tree Classifier) and a multilayer perceptron classifier (MLP Classifier) to classify a target column. Generated data’s *machine learning efficacy* measures are compared to the same measure for a train/test split of real data.

In section 5 KS Test, CS Test, Decision Tree Classifier and MLP Classifier are evaluated for:

- an 80/20 train/test split of real data
- real vs. non-distributed, and distributed discGAN generated data
- real vs. non-distributed and distributed CTGAN generated data

We propose three comparison metrics to evaluate the difference in KS Test, CS Test and machine learning efficacy between real vs. synthetic data, and real vs. a train/test split of real data. KS Test Compare, KSTC (equation 1), is used to evaluate the distance between the synthetic and real data, vs. the distance between a train/test split of the real data. CS Test Compare, CSTC (equation 2), is used to compare the probability of the synthetic and real data belongs to the same distribution vs. train/test split data belonging to the same distribution. Finally, the machine learning efficacy comparison, MLEC (equation 3), is used to evaluate the relative performance of generated data for an ML classification task, to that of a train/test split of real data.

$$\text{KSTC} = 1 - \left| \left( 1 - \frac{\text{KSTest}_{\text{generated}}}{\text{KSTest}_{\text{real}}} \right) \right| \quad (1)$$

$$\text{MLEC} = 1 - \left| \left( 1 - \frac{\text{MLE}_{\text{generated}}}{\text{MLE}_{\text{real}}} \right) \right| \quad (3)$$

$$\text{CSTC} = 1 - \left| \left( 1 - \frac{\text{CSTest}_{\text{generated}}}{\text{CSTest}_{\text{real}}} \right) \right| \quad (2)$$

In Section 5, equations 1, 2, and 3 are used to evaluate both discGAN and CTGAN synthetic records.

## 4 Preliminary Experiments

### 4.1 Proof of concept: a one-dimensional GAN

For a first proof of concept, we produced a non-conditional GAN to generate a one-dimensional dataset of eICU patient ages. The eICU dataset contains 2,500 discrete patient ages from 15 - 90, with a mean of 63.3 and standard deviation of 17.72. The original dataset combined all patient ages greater than 90 into a single category '90+'; these records were given the age of '90'. The true age distribution can be seen in orange in Figure 1.

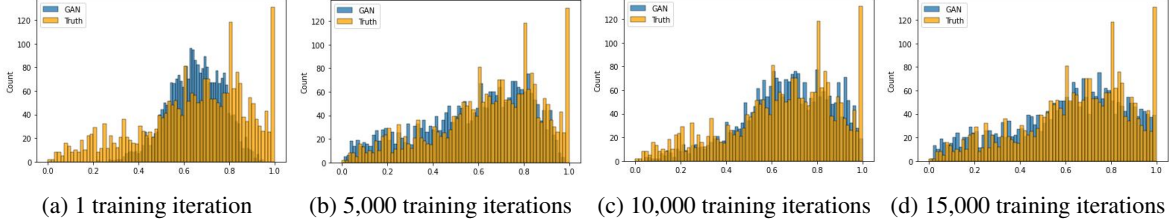


Figure 1: Real (orange) and generated (blue) age distributions while training the 1D age GAN.

A basic GAN is used to generate the one-dimensional age distribution. The discriminator input is a single value for a generated age, followed by 1) a 64-node dense layer with LeakyReLU activation ( $\alpha=0.2$ ), 2) a 10% dropout layer, 3) another 64-node dense layer with LeakyReLU activation ( $\alpha=0.2$ ), 4) a 10% dropout layer, and 5) a 1-node output layer with sigmoid activation indicating the probability of the input being real.

The generator has a random noise input of dimension 50 followed by 1) a 64-node dense layer with LeakyReLU activation ( $\alpha=0.2$ ), 2) a 64-node dense layer with LeakyReLU activation ( $\alpha=0.2$ ), and 3) a 1-node output layer with sigmoid activation. Patient age was normalized, thus, the output layer with sigmoid activation produced a synthetic age after inverse scaling.

After 15,000 iterations the results of the 1D GAN were acceptable by visual inspection of the resulting histograms of synthetically generated ages vs. real ages. The in-progress training and comparison of real and GAN-generated ages are shown in Figure 1. The mean of the generated ages was 60.2 years (compared to the real data mean of 63.3) and the standard deviation of the generated ages was 17.99 (compared to real data standard deviation of 17.72).

### 4.2 Proof of concept: a conditional tabular GAN

As our second proof of concept, we implemented a 2D conditional GAN architecture for generating 1) age-by-ethnicity for Caucasian, African American and Native American ethnicity, and 2) for age-by-unit-type for CTICU, CSICS, and Cardiac ICU unit types. Samples in our real data were significantly skewed by condition count, and had different means and standard deviations as shown in Tables 1 and 2. These results are encouraging as the generated data both quantitatively and qualitatively resembles real data; however, it is noteworthy that the standard deviation for all unit types and ethnicities is smaller in the generated data than real data, significantly so for the Native American class, which had only 12 real examples in our training data.

Table 1: Real and Generated Age by ICU Unit Type

Unit Type	Count	Real		Generated	
		Mean	Stdv	Mean	Stdv
Cardiac ICU	133	60.2	17.1	58.9	16.7
CTICU ICU	52	61.7	10.3	61.7	9.4
CSICU	65	67.6	9.3	68.3	9.3

Table 2: Real and Generated Age by Ethnicity

Ethnicity	Count	Real		Generated	
		Mean	Stdv	Mean	Stdv
African American	230	56.2	16.8	54.4	15.9
Caucasian	2010	64.4	17.4	64.4	16.1
Native American	12	50.5	19.5	50.0	9.8

A 2D discGAN uses a default noise input of 50 followed by 2 64-node dense layers with LeakyReLU activation and a dropout layer (dropout rate of 0.1). Batch sizes of 16, 32, and 64 were used, and 32 produced the most accurate results.

The following figures contain the output of the 2D GAN :

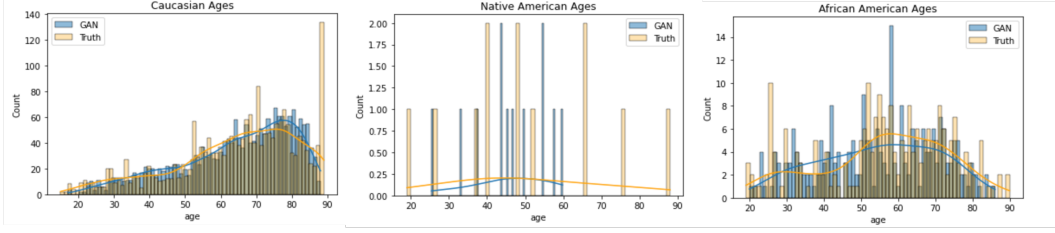


Figure 2: discGAN generated age and ethnicity data

### 4.3 Distributing a tabular GAN

discGAN models are written with the tensorflow library [14], which provides the `tf.distribute` [15] API to distribute model training across GPUs and machines. To demonstrate a tabular GAN in a distributed environment, we provisioned two GPU's and utilized `tf.distribute's MultiWorkerMirroredStrategy` which enables synchronous and distributed training using multiple GPUs on a single machine. `MultiWorkerMirroredStrategy` can distribute training across multiple workers, each with "potentially multiple GPUs." [15] More specifically, the strategy replicates and updates all model parameters, weights, and computations to each device (hence it being a "mirrored" strategy), trains on each device, then uses distributed collective implementation – namely *all-reduce*. *All-reduce* is a fused algorithm that TensorFlow uses to "aggregate tensors across all the devices by adding them up, and makes them available on each device." [15] TensorFlow uses the NVIDIA Collective Communications Library implementation of the all-reduce algorithm.

The strategy distributes discGAN by allowing the GPU workers to train together. Specifically, the devices divide the training data at each step before running forward and backward propagation separately on a mirrored copy of the the GAN models and their parameters on each device. The strategy then uses all-reduce to aggregate the distributed gradients and sends updated variable weights and a single aggregated gradient to all devices before starting the next training step.

As mentioned in 3.3, we built discGAN to distribute the generator or discriminator individually, or both together. We focused on distributing the discriminator in our experiments, with results described in 5.

Additionally, for comparative purposes, we successfully distributed CTGAN. Note that we utilized the `nn.DataParallel` class to distribute CTGAN since it is written in PyTorch rather than TensorFlow. `nn.DataParallel` follows a similar distribution implementation where models are trained on batches of training data across machines. During each training iteration, a "main" GPU broadcasts updated model parameters to each worker GPU before forward propagation is run on mini-batches of training data. Each output is sent to the main GPU, where the network loss is calculated and sent to the worker GPUs, which then run back prop. Finally, the main GPU runs all-reduce to average gradients before the updated model is sent to each worker GPU prior to the next training iteration. [16, 17]

## 5 Results

### 5.1 discGAN results

With discGAN we were able to generate multiple discrete columns in our tabular data with a single continuous column. We selected hospital discharge offset, which is the number of minutes from unit admit time that the patient was discharged from the hospital. The discrete variables included in the real data were CHF, COPD\_moderate, COPD\_nolimitations, COPD\_severe, NoHealthProblems, asthma, homeoxygen, hypertensionrequiringtreatment, restrictivepulmonarydisease, All (count of all past history entries for the patient), ethnicity, gender, and dischargestatus. discGAN is unable to generate representative data with multiple continuous features. Hospital discharge offset was selected as our continuous feature, thus the KS Test value is for hospital discharge offset in the generated data vs. real data. CS Test value is the mean chi-squared test p-value for all the discrete columns in the generated data vs. real data. Machine learning efficacy, CS Test, and KS Test values were evaluated from 249,000 generated samples from the trained discGAN model. In comparison, the real eICU data set contains 2027 records. 249,000 synthetic records were generated in order to verify discGAN can generate substantial number of statistically similar synthetic data records from a small data set.

Figure 3 shows the results of machine learning efficacy tests for past histories of congestive heart failure and severe COPD. Figures (a) and (b) show comparisons of discGAN generated data used as training data, and real data used as test

data, vs an 80/20 train/test split of real data. Figure 4 (a) shows that discGAN was able to generate joint distributions of all features with similar machine learning efficacy for congestive heart failure using a decision tree classifier. The generated data most closely matched the decision tree classifier efficacy; however, neither the real nor generated data were effective for classifying congestive heart failure.

Figure 3 (b) shows the results of the machine learning efficacy tests for the classification of a past history of severe COPD. The efficacy of the discGAN generated data vs. real data for severe COPD classification are less similar than for congestive heart failure, and interestingly, the discGAN generated data appears relatively more effective for classifying severe COPD by an MLP classifier. While more effective, this indicates the discGAN-generated data may not be statistically similar to the real data.

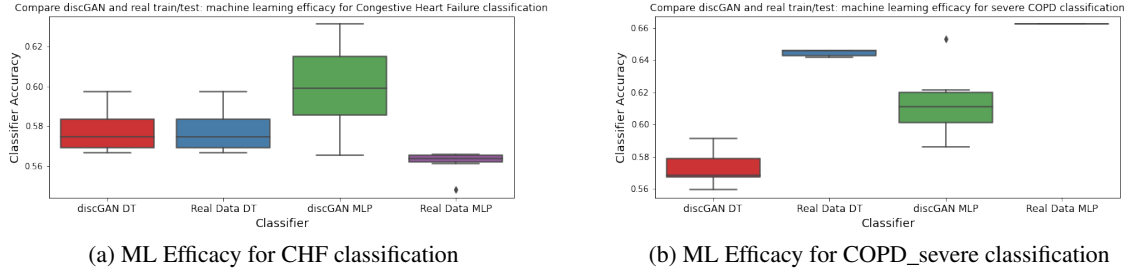


Figure 3: discGAN machine learning efficacy: (a) CHF machine learning efficacy: discGAN (red) and real (blue) data classification accuracy of Decision Tree classifiers for single-GPU and two-GPU discGANs. Results include epochs of 15,000, 20,000 and 25,000. discGAN (green) and real (purple) data classification accuracy of MLP classifiers for single-GPU and two-GPU discGANs. Results include epochs of 15,000, 20,000 and 25,000.

(b) Severe COPD machine learning efficacy: discGAN (red) and real (blue) data classification accuracy of Decision Tree classifiers for single-GPU and two-GPU discGANs. Results include epochs of 15,000, 20,000 and 25,000. discGAN (green) and real (purple) data classification accuracy of MLP classifiers for single-GPU and two-GPU discGANs. Results include epochs of 15,000, 20,000 and 25,000.

## 5.2 CTGAN results

With CTGAN we generated age, hospital discharge offset, CHF, COPD\_moderate, COPD\_nolimitations, COPD\_severe, NoHealthProblems, asthma, homeoxygen, hypertensionrequiringtreatment, restrictivepulmonarydisease, All (count of all past history entries for the patient), ethnicity, gender, and dischargestatus. Unlike with discGAN, CTGAN is able to generate multiple continuous features, in our case age and hospital discharge offset, thus the KS Test value is the average KS Test value for age and hospital discharge offset in the generated data vs. real data. Machine learning efficacy, CS Test, and KS Test values were evaluated from 5000 generated samples from the trained CTGAN model. In comparison, the real eICU data set contains 2027 records. 5000 synthetic records were generated in order to verify that a substantial number of statistically similar synthetic data records could be generated by CTGAN from an initially small data set.

Figure 4 shows the results of machine learning efficacy tests for past histories of congestive heart failure and severe COPD. Figures (a) and (b) show comparisons of CTGAN generated data used as training data, and real data used as test data, vs an 80/20 train/test split of real data. Figure 4 (a) shows CTGAN was able to generate joint distributions of features with similar machine learning efficacy for CHF. The generated data most closely matched the decision tree classifier efficacy; however, neither the real nor generated data were effective for classifying CHF. (b) shows the results of the machine learning efficacy tests for the classification of a past history of severe COPD. The efficacy of the CTGAN-generated data vs. real data for severe COPD classification are less similar than for CHF, and interestingly, the CTGAN generated data appears relatively more effective for classifying severe COPD by an MLP classifier. While more effective, similarly to discGAN, this indicates the CTGAN-generated data may also not be statistically similar to the real data.



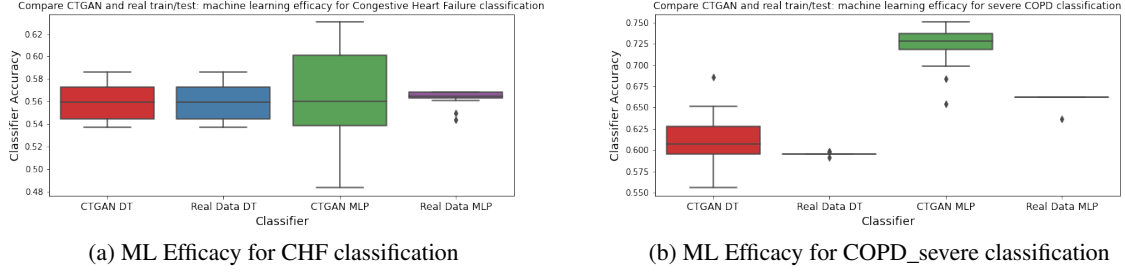


Figure 4: CTGAN machine learning efficacy: (a) CHF machine learning efficacy: CTGAN (red) and real (blue) data classification accuracy of Decision Tree classifiers for single-GPU and two-GPU CTGANs. Results include epochs of 15,000, 20,000 and 25,000. CTGAN (green) and real (purple) data classification accuracy of MLP classifiers for single-GPU and two-GPU CTGANs. Results include epochs of 15,000, 20,000 and 25,000.

(b) Severe COPD machine learning efficacy: CTGAN (red) and real (blue) data classification accuracy of Decision Tree classifiers for single-GPU and two-GPU CTGANs. Results include epochs of 15,000, 20,000 and 25,000. CTGAN (green) and real (purple) data classification accuracy of MLP classifiers for single-GPU and two-GPU CTGANs. Results include epochs of 15,000, 20,000 and 25,000.

### 5.3 Comparison: single GPU vs. two GPUs

For both discGAN and CTGAN we evaluated KS Test and CS Test values by training epochs. Unlike common evaluation tools for image GANs, this comparison provides a quantitative measure of model performance by training time; KS Test and CS Test values were also compared by a GAN trained using a single GPU, and the GAN trained on two GPUs. Figure 5 displays KS Test and CS Test values for 249,000 discGAN generated synthetic records using 15,000, 20,000, and 25,000 epochs and compares single vs. multi-GPU test values. Figure 5 displays KS Test and CS Test values for 5,000 CTGAN-generated synthetic records using 1,000, 5,000, 10,000, 15,000, 20,000, and 25,000 epochs and compares single vs. multi-GPU test values.

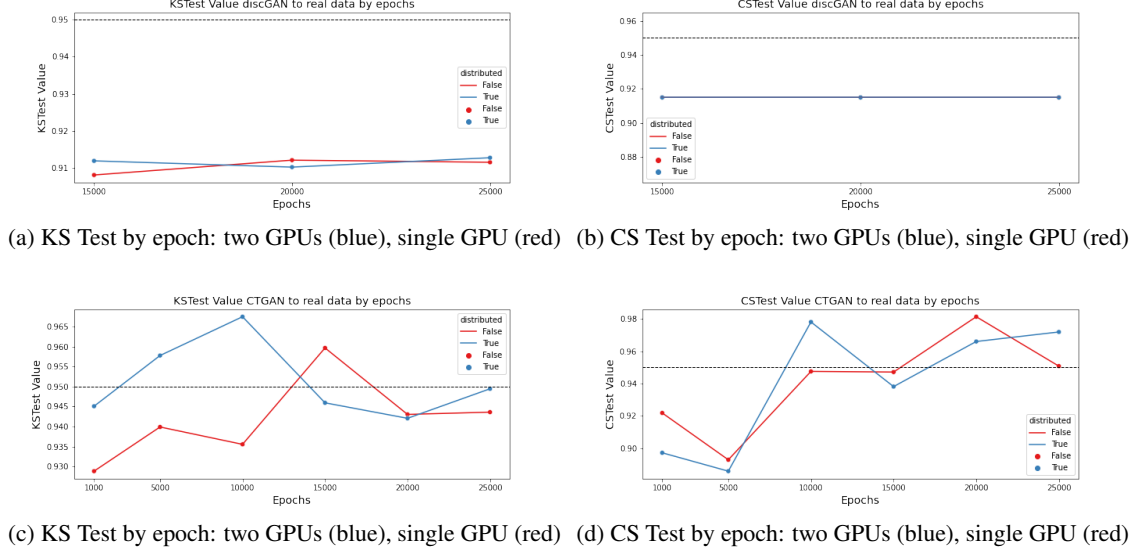


Figure 5: (a) and (b) discGAN: KS Test and CS Test values by training epoch for discGAN trained on a single GPU vs. two GPUs evaluated on 249,000 synthetic data rows.

(c) and (d) CTGAN KS Test and CS Test values by training epoch for discGAN trained on a single GPU vs. two GPUs evaluated on 5,000 synthetic data rows.

For CTGAN, we evaluated the time-per-epoch training when the GAN was trained on a single GPU and two GPUs. Results are shown in Figure 6 and discussed in Section 6.



Figure 6: Seconds per epoch training CTGAN on a single GPU and two GPUs.

#### 5.4 discGAN vs. CTGAN

	Single GPU	2-GPU
Evaluation Criteria		
KS Test	0.911	0.913
KSTC	0.922	0.923
CS Test	0.915	0.915
CSTC	0.953	0.953
MLEC Decision Tree Congestive Heart Failure	1.0	1.0
MLEC MLP Congestive Heart Failure	0.999	0.875
MLEC Decision Tree Severe COPD	0.9	0.916
MLEC MLP Tree Severe COPD	0.939	0.904

Table 3: discGAN evaluation results for 249,000 synthetic records after 25,000 training epochs. Single-GPU and 2-GPU training are compared.

	Single GPU	2-GPU
Evaluation Criteria		
KS Test	0.944	0.949
KSTC	0.956	0.962
CS Test	0.951	0.972
CSTC	0.99	0.988
MLEC Decision Tree Congestive Heart Failure	1.0	1.0
MLEC MLP Congestive Heart Failure	0.996	0.975
MLEC Decision Tree Severe COPD	0.988	0.976
MLEC MLP Tree Severe COPD	0.898	0.868

Table 4: CTGAN evaluation results for 5,000 synthetic records after 25,000 training epochs. Single-GPU and 2-GPU training are compared.

## 6 GitHub Repository

Repository contains all code related to generating and evaluating healthcare data using discGAN as well as files associated with distributing CTGAN. <https://github.com/dianamJLAB/DS6050>

## 7 Discussion and Conclusion

We built a distributed, conditional tabular GAN, discGAN, to produce synthetic tabular health-care data. Our architecture can be used to handle many discrete features and a single continuous feature, with distribution possible for both the

generator and discriminator together or individually. We hope to better generalize our GAN architecture to handle any number of continuous features in future work.

We performed several experiments using discGAN in which we generated privatized synthetic data, evaluating performance with both qualitative distribution analysis and quantitative methods, including machine learning efficacy, and Kolmogorov-Smirnov and Chi-squared tests. For discrete features such as congestive heart failure, discGAN created distributions similar to the real data. For other discrete features such as severe COPD, discGAN generated data was not similar to the real data, possibly because of noise in the real data due to inherent subjectivity in labeling COPD level. For KS Test and CS Test, discGAN performed similarly to the SOTA CTGAN, but CTGAN was able to generate multiple continuous features, unlike discGAN.

Distribution of discGAN from a single GPU to two GPUs had limited-to-no effect on training time or our performance measures, including KS Test, CS Test, and machine learning efficacy. The lack of improvement in training time is likely due to the small size of the eICU dataset, resulting in any GPU acceleration of training being canceled out with the time it takes to communicate between the distributed workers. However, the motivation for a distributed, tabular healthcare GAN is not solely to reduce training time; more importantly, its purpose is to allow privacy of real data across multiple workers.

Interestingly, when we updated the CTGAN source code so that we can distribute its training across GPUs, the distributed CTGAN was able to obtain higher KS Test and CS Test values in fewer epochs than a single-GPU CTGAN. Figure 5 demonstrates that for CTGAN’s generation of 5000 synthetic data records, the distributed CTGAN had the highest KS Test value, and it achieved this value in fewer epochs (10,000 epochs) than the highest KS Test value from the CTGAN trained on the single GPU (15,000 epochs). Similarly, the distributed model obtained its highest CS Test value earlier (10,000 epochs) than the single-GPU-trained CTGAN model (20,000 epochs). This demonstrates the effectiveness of distributed training for a GAN.

## References

- [1] Johnson A. Raffa J. Celi L. A. Badawi O. Mark R. Pollard, T. eicu collaborative research database (version 2.0). *PhysioNet*, 2019.
- [2] Tom J. Pollard, Alistair E. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1), 2018.
- [3] Ary L. Goldberger, Luis A. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, H. Eugene Stanley, and et al. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23), 2000.
- [4] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, 2019.
- [5] N. Patki, R. Wedge, and K. Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, Oct 2016.
- [6] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U. Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Advances in Neural Information Processing Systems, pages 3308–3318. Neural Information Processing Systems, December 2017. NIPS 2017 : 31st Conference on Neural Information Processing Systems, NIPS 2017 ; Conference date: 04-12-2017 Through 09-12-2017.
- [7] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083, 2018. Publisher Copyright: © 2018 VLDB Endowment 21508097/18/4.; 44th International Conference on Very Large Data Bases, VLDB 2018 ; Conference date: 27-08-2018 Through 31-08-2018.
- [8] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR, 18–19 Aug 2017.
- [9] Corentin Hardy, Erwan Le Merrer, and Bruno Sericola. Md-gan: Multi-discriminator generative adversarial networks for distributed datasets. In *2019 IEEE international parallel and distributed processing symposium (IPDPS)*, pages 866–877. IEEE, 2019.

- [10] Rachid Guerraoui, Arsany Guirguis, Anne-Marie Kermarrec, and Erwan Le Merrer. Fegan: Scaling distributed gans. In *Proceedings of the 21st International Middleware Conference*, Middleware '20, page 193–206, New York, NY, USA, 2020. Association for Computing Machinery.
- [11] Qi Chang, Hui Qu, Yikai Zhang, Mert Sabuncu, Chao Chen, Tong Zhang, and Dimitris N Metaxas. Synthetic learning: Learn from distributed asynchronized discriminator gan without sharing medical image data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13856–13866, 2020.
- [12] Nist/sematech e-handbook of statistical methods, Apr 2012.
- [13] Stavroula Bourou, Andreas El Saer, Terpsichori-Helen Velivasaki, Artemis Voulkidis, and Theodore Zahariadis. A review of tabular data synthesis using gans on an ids dataset. *Information*, 12:375, 09 2021.
- [14] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [15] Distributed training with tensorflow : Tensorflow core.
- [16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [17] Ankur Mohan. Distributed data parallel training using pytorch on aws. <https://www.telesens.co/2019/04/04/distributed-data-parallel-training-using-pytorch-on-aws/>, Apr 2019.