Knowledge Intensive Learning of Generative Adversarial Networks

Devendra Singh Dhami
devendra.dhami@utdallas.edu
The University of Texas at Dallas

Mayukh Das
Samsung Research India
mayukh.das@samsung.com

Sriraam Natarajan
The University of Texas at Dallas
sriraam.natarajan@utdallas.edu

ABSTRACT

While Generative Adversarial Networks (GANs) have accelerated the use of generative modelling within the machine learning community, most of the applications of GANs are restricted to images. The use of GANs to generate clinical data has been rare due to the inability of GANs to faithfully capture the intrinsic relationships between features. We hypothesize and verify that this challenge can be mitigated by incorporating domain knowledge in the generative process. Specifically, we propose human-allied GANs that using correlation advice from humans to create synthetic clinical data. Our empirical evaluation demonstrates the superiority of our approach over other GAN models.

CCS CONCEPTS

• **Deep Learning** → **Generative Adversarial Networks**; • **Application** → *Healthcare*; • **Learning** → *Knowledge Intensive Learning*.

KEYWORDS

generative adversarial networks, human in the loop, healthcare

ACM Reference Format:
Devendra Singh Dhami, Mayukh Das, and Sriraam Natarajan. 2020. Knowledge Intensive Learning of Generative Adversarial Networks. In *San Diego '20: Knowledge-infused Mining and Learning, 2020, San Diego, CA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Deep learning models have reshaped the machine learning landscape over the past decade [16, 29]. Specifically, Generative Adversarial Networks (GANs) [17] have found tremendous success in generating examples for images [34, 37, 45], photographs of human faces [1, 25, 52], image to image translation [30, 33, 55] and 3D object generation [44, 51, 53] to name a few. Despite such success, there are several key factors that limit the widespread adoption of GANs, for a broader range of tasks, including, widely acknowledged data hungry nature of such methods, potential access issues of real medical data and finally, their restricted usage, mainly in the context of images. These factors have limited the use of these arguably successful techniques in medical (or similar) domains. However, recently, synthetic data generation has become a centerpiece of research in medical AI due to the diverse difficulties in collection, persistence, sharing and analysis of real clinical data.

We aim to address the above limitations. Inspired by Mitchell's argument of "The Need for Biases in Learning Generalizations" [38], we mitigate the challenges of existing data hungry methods via inductive bias while learning GANs. We show that effective inductive bias can be provided by humans in the form of domain knowledge [14, 27, 41, 50]. Rich human advice can effectively balance the impact of quality (sparsity) of training data. Data quality also contributes to, the well studied, modal instability of GANs. This problem is especially critical in domains such as medical/clinical analytics that does not typically exhibit 'spatial homophily' [21], unlike images, and are prone to distributional diversity among feature clusters as well. Our human-guided framework proposes a robust strategy to address this challenge. Note that in our setting the human is an ally and not an adversary.

The second limitation of access is crucial for medical data generation. Access to existing medical databases [10, 18] is hard due to cost and access concerns and thus synthetic data generation holds tremendous promise [6, 13, 19, 35, 48]. While previous methods generated synthetic images, we go beyond images and generate clinical data. Building on this body of work, we present a synthetic data generation framework that effectively exploits domain expertise to handle data quality.

We make a few key contributions:

(1) We demonstrate how effective human advice can be provided to a GAN as an inductive bias.
(2) We present a method for generating data given this advice.
(3) Finally, we demonstrate the effectiveness and efficacy of our approach on 2 de-identified clinical data sets. Our method is generalizable to multiple modalities of data and is not necessarily restricted to images.
(4) Yet another feature of this approach is that training occurs from very few data samples (<50 in one domain) thus providing human guidance as a data generation alternative.

2 RELATED WORK

The key principle behind GANs [17] is a zero-sum game [26] from game theory, a mathematical representation where each participant's gain or loss is exactly balanced by the losses or gains of the other participants and is generally solved by a minimax algorithm. The generator distribution $p_{data}(x)$ over the given data x is learned by sampling z from a random distribution $p_z(z)$ (initially uniform was proposed but Gaussians have been proven superior [2]). While GANs have proven to be a powerful framework for estimating generative distributions, convergence dynamics of naive mini-max algorithm has been shown to be unstable. Some recent approaches, among

San Diego '20, 2020, San Diego, CA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

many others, augment learning either via statistical relationships between true and learned generative distributions such as Wasserstein-1 distance [3], MMD [32] or via spectral normalization of the parameter space of the generator [39] which controls the generator distribution from drifting too far. Although these approaches have improved the GAN learning in some cases, there is room for improvement.

Guidance via human knowledge is a provably effective way to control learning in presence of systematic noise (which leads to instability). One typical strategy to incorporate such guidance is by providing *rules over training examples and features*. Some of the earliest approaches are explanation-based learning (EBL-NN, [49]) or ANNs augmented with symbolic rules (KBANN, [50]). Various widely-studied techniques of leveraging domain knowledge for optimal model generalization include polyhedral constraints in case of knowledge-based SVMs, [9, 14, 28, 47]), preferences rules [5, 27, 41, 42] or qualitative constraints (ex: monotonicities / synergies [54] or quantitative relationships [15]). Notably, whereas these models exhibit considerable improvement with the incorporation of human knowledge, there is only limited use of such knowledge in training GANs. Our approach resembles the qualitative constraints framework in spirit.

While widely successful in building optimally generalized models in presence of systematic noise (or sample biases), knowledge-based approaches have mostly been explored in the context of discriminative modeling. In the generative setting, a recent work extends the principle of posterior regularization from Bayesian modeling to deep generative models in order to incorporate structured domain knowledge [22]. Traditionally, knowledge based generative learning has been studied as a part of learning probabilistic graphical models with structure/parameter priors [36]. We aim to extend the use of knowledge to the generative model setting.

3 KNOWLEDGE INTENSIVE LEARNING OF GENERATIVE ADVERSARIAL NETWORKS

A notable disadvantage of adversarial training formulation is that the training is slow and unstable, leading to mode collapse [2] where the generator starts generating data of only a single modality. This has resulted in GANs not being exploited to their full potential in generating synthetic non-image clinical data. Human advice can encourage exploration in diverse areas of the feature space and helps learn more stable models [43]. Hence, we propose a human-allied GAN architecture (HA-GAN) (figure 1). The architecture incorporates human advice in form of feature correlations. Such intrinsic relationships between the features are crucial in medical data sets and thus become a natural candidate as additional knowledge/advice in guided model learning for faithful data generation.

Our approach builds upon a GAN architecture [17] where a random noise vector is provided to the generator which tries to generate examples as close to the real distribution as possible. The discriminator tries to distinguish between real examples and ones generated by the generator. The generator tries to maximize the probability that the discriminator makes a mistake and the discriminator tries to minimize its mistakes thereby resulting in a min-max optimization problem which can be solved by a mini-max algorithm. We adopt the Wasserstein GAN (WGAN) architecture[1] [3, 20] that focuses on defining a distance/divergence (Wasserstein or earth movers distance) to measure the closeness between the real distribution and the model distribution.

3.1 Human input as inductive bias

Historically, two approaches have been studied for using guidance as bias. The **first** is to provide advice on the labels as constraints or preferences that controls the search space. Some example advice rules on the labels include: *$(3 \leq feature_1 \leq 5) \Rightarrow label = 1$* and *$(0.6 \leq feature_2 \leq 0.8) \wedge (4 \leq feature_3 \leq 5) \Rightarrow label = 0$*. Such advice is more relevant in an discriminative setting but are not ideal for GANs. Since GANs are shown to be sensitive to the training data and here the *labels are getting generated*, they should not be altered during training. The **second** is via correlations between features as preferences (*our approach*) which allows for faithful representation of diverse modality.

Advice injection: After every fixed number of iterations, N, we calculate the correlation matrix of the generated data $\mathcal{G}_1$ and provide a set of advice ψ on the correlations between different features. Consider the following motivating example for the use of correlations as a form of advice.

Example: *Consider predicting heart attack with 3 features - cholesterol, blood pressure (BP) and income. The values of the given features can vary (sometimes widely) between different patients due to several latent factors (ex, smoking habits). It is difficult to assume any specific distribution. In other words, it is difficult to deduce whether the values for the features come from the same distribution (even though the feature values in the data set are similar).*

We modify the correlation coefficients (for both positive and negative correlations) between the features by increasing them if the human advice suggests that two features are highly correlated and decrease the same if the advice suggests otherwise.

Example: *Continuing the above example, since rise in the cholesterol level can lead to rise in BP and vice versa, expert advice here can suggest that cholesterol and BP should be highly correlated. Also, as income may not contribute directly to BP and cholesterol levels, another advice here can be to de-correlate cholesterol/BP and income level.*

The example advice rules $\in \psi$ are: 1. Correlation("cholesterol level","BP")↑, 2. Correlation("cholesterol level","income level")↓ and 3. Correlation("BP","income level")↓, where ↑ and ↓ indicate increase and decrease respectively. Based on the 1st advice we need to increase the correlation coefficient between *cholesterol level* and *BP*. Then

$$C = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \mathcal{A} = \begin{bmatrix} 1 & \lambda & 1 \\ \lambda & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \qquad (1)$$

Here C is the correlation matrix, $\mathcal{A}$ is the advice matrix and λ is the factor by which the correlation value is to be augmented. In case where we need to increase the value of the correlation coefficient, λ should be > 1. We keep $\lambda = \frac{1}{max(|C|)}$. Since $-1.0 \leq \forall c \in C \leq 1.0$, in this case, the value of $\lambda \geq 1.0$, leading to enhanced correlation via

[1] We use 'GAN' to indicate 'W-GAN'

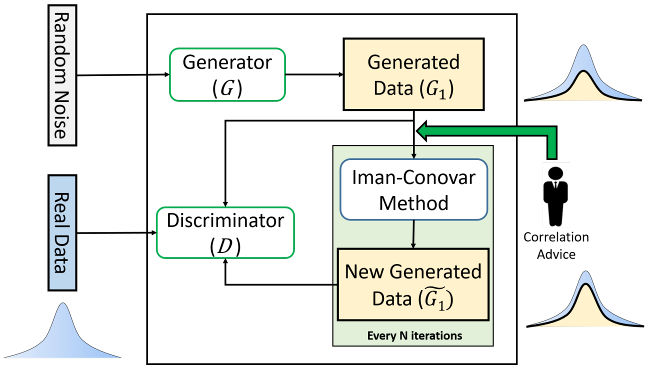

Figure 1: Human-Allied GAN. Correlation advice takes generated distribution closer to the real distribution.

Hadamard product. Thus the new correlation matrix $\hat{C}$ is,

$$\begin{aligned}\hat{C} = C \odot \mathcal{A} &= \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 & \frac{1}{0.3} & 1 \\ \frac{1}{0.3} & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0.667 & 0.3 \\ 0.667 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix}\end{aligned} \tag{2}$$

If the advice says that features have low correlations (2nd rule in example), we decrease the correlation coefficient. Now, λ must be < 1 and we set $\lambda = max(|C|)$. Since -1 $\leq \forall c \in C \leq 1.0$, the value of $\lambda \leq 1.0$. Thus multiplying by λ will decrease the correlation value, and the new correlation matrix is,

$$\begin{aligned}\hat{C}_1 = \hat{C} \odot \mathcal{A} &= \begin{bmatrix} 1 & 0.667 & 0.3 \\ 0.667 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 & 1 & 0.3 \\ 1 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0.667 & 0.09 \\ 0.667 & 1 & 0.021 \\ 0.09 & 0.021 & 1 \end{bmatrix}\end{aligned} \tag{3}$$

This is used to create the new generated data $\tilde{\mathcal{G}}_1$. For negative correlations, the process is unchanged.

3.2 Advice-guided data generation

After $\hat{C}_1$ is constructed, we next generate data satisfying the constraints. To this effect, we employ the Iman-Conover method [23], a distribution free method to define dependencies between distributional variables based on rank correlations such as Spearman or Kendell Tau correlations. Since we deal with linear relationships between the features and assume a normal distribution and that Pearson coefficient has shown to perform equally well with the Iman-Conover method [40] due to the close relationship between Pearson and Spearman correlations, we use the Pearson correlations. Further, we assume that the features are Gaussian, justified by the fact that most lab test data is continuous. The Iman-Conover method consists of the following steps:

[Step 1]: Create a random standardized matrix $\mathcal{M}$ with values $x \in \mathcal{M} \sim$ Gaussian distribution. This is obtained by the process of inverse transform sampling described next. Let $\mathcal{V}_1$ be a uniformly distributed random variable and $\mathcal{CDF}$ be the cumulative distribution function. For a sampled point v, $\mathcal{CDF}(v) = \mathcal{P}(V \leq v)$. Thus, to generate samples, the values $v \sim \mathcal{V}$ are passed through $\mathcal{CDF}^{-1}$ to obtain the desired values x $[\mathcal{CDF}^{-1}(v) = \{x | \mathcal{CDF}(x) \leq v, v \in [0, 1]\}]$. Thus for Gaussian,

$$\begin{aligned}\mathcal{CDF}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp^{\frac{-x^2}{2}} dx &= \frac{1}{\sqrt{2\pi}} \int_{0}^{x} \exp^{\frac{-x^2}{2}} dx \\ &= [-\exp(\frac{-x^2}{2})]_0^x\end{aligned} \tag{4}$$

The inverse CDF can be thus written as $\mathcal{CDF}^{-1}(v) = 1{-}\exp(\frac{-x^2}{2}) \leq v$ and the desired values $x \in \mathcal{M}$ can be obtained as $x = \sqrt{2ln(1-v)}$.

[Step 2]: Calculate the correlation matrix $\mathcal{E}$ of $\mathcal{M}$.

[Step 3]: Calculate the Cholesky decomposition $\mathcal{F}$ of the correlation matrix $\mathcal{E}$. Cholesky decomposition [46] of a *positive-definite* matrix is given as the product of a lower triangular matrix and its conjugate transpose. Note that for Cholesky decomposition to be unique, the target matrix should be positive definite, (such as the co-variance matrix) whereas the correlation matrix, used in our algorithm, is only positive semi-definite. We enforce positive-definiteness by repeated addition of very small values to the diagonal of the correlation matrix until positive-definiteness is ensured. Given a symmetric and positive definite matrix $\mathcal{E}$, its Cholesky decomposition $\mathcal{F}$ is such that $\mathcal{E} = \mathcal{F} \cdot \mathcal{F}^{\top}$.

[Step 4]: Calculate the Cholesky decomposition Q of the correlation matrix obtained after modifications based on human advice, $\hat{C}$. As above the Cholesky decomposition is such that $\hat{C} = Q \cdot Q^{\top}$.

[Step 5]: Calculate the reference matrix $\mathcal{T}$ by transforming the sampled matrix $\mathcal{M}$ from step 1 to have the desired correlations of $\hat{C}$, by using their Cholesky decompositions.

[Step 6]: Rearrange values in columns of the generated data $\mathcal{G}_1$ to have the same ordering as corrresponding column in the reference matrix $\mathcal{T}$ to obtain the final generated data $\tilde{\mathcal{G}}_1$.

Cholesky decomposition to model correlations: Given an randomly generated data set with no correlations $\mathcal{P}$, a correlation matrix C and its Cholesky decomposition Q, data that faithfully follows the given correlations $\in C$ can be generated by the product of the obtained lower triangular matrix with the original uncorrelated data i.e. $\hat{\mathcal{P}}=Q\mathcal{P}$. The correlation of the newly obtained data, $\hat{\mathcal{P}}$ is,

$$Corr(\hat{\mathcal{P}}) = \frac{Cov(\hat{\mathcal{P}})}{\sigma_{\hat{\mathcal{P}}}} = \frac{\mathrm{E}[\hat{\mathcal{P}}\hat{\mathcal{P}}^{\top}] - \mathrm{E}[\hat{\mathcal{P}}]\mathrm{E}[\hat{\mathcal{P}}]^{\top}}{\sigma_{\hat{\mathcal{P}}}} \tag{5}$$

Since we consider data $\hat{\mathcal{P}}$ from a Gaussian distribution with zero mean and unit variance,

$$\begin{aligned} Corr(\hat{\mathcal{P}}) &= \frac{\mathrm{E}[\hat{\mathcal{P}}\hat{\mathcal{P}}^{\top}] - \mathrm{E}[\hat{\mathcal{P}}]\mathrm{E}[\hat{\mathcal{P}}]^{\top}}{\sigma_{\hat{\mathcal{P}}}} = \mathrm{E}[\hat{\mathcal{P}}\hat{\mathcal{P}}^{\top}] = \mathrm{E}[(Q\mathcal{P})(Q\mathcal{P})^{\top}] \\ &= \mathrm{E}[Q\mathcal{P}Q^{\top}\mathcal{P}^{\top}] = Q\mathrm{E}[\mathcal{P}\mathcal{P}^{\top}]Q^{\top} = QQ^{\top} = C \end{aligned} \tag{6}$$

Thus Cholesky decomposition can capture the desired correlations faithfully and can be used for generating correlated data. Since we already have a normal sampled matrix $\mathcal{M}$ and a calculated correlation $\mathcal{E}$ of $\mathcal{M}$, we need to calculate a reference matrix (step 5).

3.3 Human-Allied GAN training

Since the human expert advice is provided independent of the GAN architecture, our method is agnostic of the underlying GAN architecture. We make use of Wasserstein GAN (WGAN) architecture since its shown to be more stable while training and can handle mode collapse [3]. Only the error backpropagation values differ when we are using the data generated by the underlying GAN or the data generated by the Iman-Conover method. Our algorithm starts with the general process of training a GAN where the generator takes random noise as an input and generates data which is then passed, along with the real data, to the discriminator. The discriminator tries to identify the real and generated data and the error is back propagated to the generator. After every specified number of iterations, the correlations between features C in the generated data is obtained and a new correlation matrix $\hat{C}$, is obtained with respect to the expert advice (section 3.1). A new data set is generated wrt $\hat{C}$ using the Iman-Conover method (Section 3.2) and then passed to the discriminator along with the real data set.

4 EXPERIMENTAL EVALUATION

We aim to answer the following questions:

- **Q1:** Does providing advice to GANs help in generating better quality data?
- **Q2:** Are GANs with advice effective for data sets that have few examples?
- **Q3:** How does bad advice affect the quality of generated data?
- **Q4:** How well does human advice handle class imbalance?
- **Q5:** How does our method compare to state-of-the-art GAN architectures.

We consider 2 *real clinical data sets*.

(1) **Nephrotic Syndrome** is a novel data set of symptoms that indicate kidney damage. This consists of 50 kidney biopsy images along with the clinical reports sourced from Dr Lal PathLabs, India [2]. We use the clinical reports that consist of the values for kidney tissue diagnosis which can confirm the clinical diagnosis and help to identify high-risk patients and influence treatment decisions and help medical practitioners to plan and prognosticate treatments. The data consists of 19 features with 44 positive and 6 negative examples.

(2) **MIMIC** database [24] consists of deidentified information of patients admitted to critical care units at a large tertiary care hospital. The features included are predominately time window aggregations of physiological measurements from the medical records. We selected relevant lab results, vital sign observations and feature aggregations. The data consists of 18 with 5813 positive and 40707 negative examples.

Advice Acquisition: Here we compile the sources from which we obtain the advice.

(1) *Nephrotic Syndrome:* This is a novel real data set and the **advice is obtained from a nephrologist in India**. According to the problem statement from the expert, nephrotic syndrome involves the loss of a lot of protein and nephritic syndrome involves the loss of a lot of blood through urine. A kidney biopsy is often required to diagnose the underlying pathology in patients with suspected glomerular disease. The goal of the project is to build a clinical support system that predicts the disease using clinical features, thus reducing the need of kidney biopsy. ***Since the data collection is scarce, a synthetic data set can help in better understanding of the disease from the clinical features.***

(2) *MIMIC:* The feature set and the expected correlations are obtained in consultation with **trauma experts at a Dallas hospital**.

All experiments were run on a *64-bit Intel(R) Xeon(R) CPU E5-2630 v3* server for 10K epochs. Both the generator and discriminator are neural networks with 4 hidden layers. To measure the quality of the generated data we make use of the *train on synthetic, test on real* (TSTR) method as proposed in [12]. We use gradient boosting with 100 estimators and a learning rate of 0.01 as the underlying model. We train the GAN for 10K epochs and provide correlation advice every 1K iterations.

Table 1 shows the results of the TSTR method with data generated with (HA-GAN$_{GA}$) and without advice (GAN). It shows that the data generated with advice has higher TSTR performance than the data generated without advice across all data sets and all metrics. Thus, to answer **Q1**, providing advice to generative adversarial networks captures the relationship between features better and thus are able to generate better quality synthetic data.

Learning with less data: GANs with advice are especially impressive in nephrotic syndrome data which consists of only 50 examples across all metrics and is thus very small in size when compared to the number of samples typically required to train a GAN model. Thus, we realize an important property of incorporating human guidance in the GAN model and can answer **Q2** affirmatively. *The use of advice opens up the potential of using GANs in presence of sparse data samples.*

Effect of bad advice: Table 1 also shows the results for data generated with bad advice (HA-GAN$_{BA}$). To simulate bad advice, we follow a simple process: if the advice says that the correlation between features should be high, we set the correlations in $\hat{C}$ to 0 and if the advice says that the correlation should be low, we set the correlations in $\hat{C}$ to be either 1 or -1 based on whether the original

[2] https://www.lalpathlabs.com/

Table 1: TSTR Results ($\approx 3\ dec.$). N/A in Nephrotic Syndrome denotes that all generated labels were of a single class (0 in our case) and thus we were not able to run the discriminative algorithm in the TSTR method. GA **and** BA **denotes good and bad advice to our HA-GAN model respectively.**

Data set	Methods	Recall	F1	AUC-ROC	AUC-PR
NS	GAN	0.584	0.666	0.509	0.911
	HA-GAN$_{BA}$	0.42	0.511	0.518	0.886
	medGAN	N/A	N/A	N/A	N/A
	medWGAN	N/A	N/A	N/A	N/A
	medBGAN	N/A	N/A	N/A	N/A
	HA-GAN$_{GA}$	**1.0**	**0.943**	**0.566**	**0.947**
MIMIC	GAN	0.122	0.119	0.495	0.174
	HA-GAN$_{BA}$	0.285	0.143	0.459	0.235
	medGAN	0.374	0.163	0.478	0.279
	medWGAN	0.0	0.0	0.5	0.562
	medBGAN	0.0	0.0	0.5	0.562
	HA-GAN$_{GA}$	**0.979**	**0.263**	**0.598**	**0.567**

correlation is positive or negative. Thus, given a correlation matrix

$$C = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \tag{7}$$

suppose the advice says that we need to increase the correlation coefficient between feature 1 and feature 2. Then the new correlation matrix after bad advice can be calculated as:

$$C = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \mathcal{A} = \begin{bmatrix} 1 & \lambda & 1 \\ \lambda & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{8}$$

$$\hat{C} = C \odot \mathcal{A} = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 & \lambda & 1 \\ \lambda & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{9}$$

where λ is the factor by which the correlation value is to be augmented. Since the advice asks to increase the correlation, we set λ=0. Thus,

$$\hat{C} = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.0 & 0.3 \\ 0.0 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \tag{10}$$

Similarly, if the advice says that we need to decrease the correlation coefficient between feature 1 and feature 3, we set $\lambda = \frac{1}{feat_{val}}$.

$$\hat{C} = \begin{bmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.07 \\ 0.3 & 0.07 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 & 0.2 & \frac{1}{0.3} \\ 0.2 & 1 & 1 \\ \frac{1}{0.3} & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.2 & 1.0 \\ 0.2 & 1 & 0.07 \\ 1.0 & 0.07 & 1 \end{bmatrix} \tag{11}$$

As results show in table 1, giving bad advice adversely affects the performance thereby answering **Q3**.

The nephrotic syndrome and MIMIC data sets are relatively unbalanced with a pos to neg ratio of ≈ 8:1 and 1:7 respectively. Most of the medical data sets, except highly curated data sets, are unbalanced. A data generator model should be able to handle this imbalance. Since our method explicitly focuses on the correlations between features and generates better quality data based on such relationships between features, our method is quite **robust to the imbalance in the underlying data**. This can be seen in the results in table 1 where advice based data generation outperforms the non-advice and bad advice based data generation. Thus, we can answer **Q4** affirmatively.

To answer **Q5** we compare our method to 3 GAN architectures, medGAN [8] which uses an encoder decoder framework for EHR data generation and its 2 variants medBGAN and medWGAN [4] and the results are shown in table 1. Our method, with good advice, outperforms the baseline both domains showing the effectiveness of our method.

5 CONCLUSION

We presented a new GAN formulation that employs correlation information between features as advice to generate new correlated data and train the underlying GAN model. We tested our model on real clinical data sets and show that incorporating advice helps generate good quality synthetic medical data. We employ TSTR method to test the quality of generated data and demonstrated that the generated data with advice is more aligned with the real data. There are several future interesting directions. First, providing advice only when required in an active fashion can allow for significant reduction in the amount of effort on the human side. Second, there can be multiple advice options, such as posterior regularization [15], that can be used to capture feature relationships explicitly. Third, although we do not have identifiers in the data, thereby eliminating the need of differential privacy [11], a general framework that can uphold the privacy of patient data along the lines of using Cholesky decomposition [7, 31] is a natural next step.

ACKNOWLEDGMENTS

DSD and SN gratefully acknowledge DARPA Minerva award FA9550-19-1-0391. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA or the US government.

REFERENCES

[1] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. 2017. Face aging with conditional generative adversarial networks. In *ICIP*.
[2] Martin Arjovsky and Leon Bottou. 2017. Towards principled methods for training generative adversarial networks. In *ICLR*.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *ICML* (2017).

[4] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. 2019. Synthesizing electronic health records using improved generative adversarial networks. *JAMA* (2019).

[5] Darius Braziunas and Craig Boutilier. 2006. Preference elicitation and generalized additive utility. In *AAAI*.

[6] Anna L Buczak, Steven Babin, and Linda Moniz. 2010. Data-driven approach for creating synthetic electronic medical records. *BMC medical informatics and decision making* (2010).

[7] Jim Burridge. 2003. Information preserving statistical obfuscation. *Statistics and Computing* (2003).

[8] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. 2017. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. In *MLHC*.

[9] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* (1995).

[10] Ivo D Dinov. 2016. Volume and value of big healthcare data. *Journal of medical statistics and informatics* (2016).

[11] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *TAMS*.

[12] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. 2017. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633* (2017).

[13] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. In *ISBI*.

[14] Glenn M Fung, Olvi L Mangasarian, and Jude W Shavlik. 2003. Knowledge-based support vector machine classifiers. In *NIPS*.

[15] Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. 2010. Posterior regularization for structured latent variable models. *JMLR* (2010).

[16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.

[18] Peter Groves, Basel Kayyali, David Knott, and Steve Van Kuiken. 2016. The'big data'revolution in healthcare: Accelerating value and innovation. (2016).

[19] John T Guibas, Tejpal S Virdi, and Peter S Li. 2017. Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872* (2017).

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *NIPS*.

[21] Haroun Habeeb, Ankit Anand, Mausam Mausam, and Parag Singla. 2017. Coarse-to-fine lifted MAP inference in computer vision. In *IJCAI*.

[22] Zhiting Hu, Zichao Yang, Russ R Salakhutdinov, LIANHUI Qin, Xiaodan Liang, Haoye Dong, and Eric P Xing. 2018. Deep Generative Models with Learnable Knowledge Constraints. In *NeurIPS*.

[23] Ronald L Iman and William-Jay Conover. 1982. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics-Simulation and Computation* (1982).

[24] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* (2016).

[25] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*.

[26] Harold William Kuhn and Albert William Tucker. 1953. *Contributions to the Theory of Games*.

[27] Gautam Kunapuli, Phillip Odom, Jude W Shavlik, and Sriraam Natarajan. 2013. Guiding autonomous agents to better behaviors through human advice. In *ICDM*.

[28] Quoc V Le, Alex J Smola, and Thomas Gärtner. 2006. Simpler knowledge-based support vector machines. In *ICML*.

[29] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* (2015).

[30] Minjun Li, Haozhi Huang, Lin Ma, Wei Liu, Tong Zhang, and Yugang Jiang. 2018. Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks. In *ECCV*.

[31] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang. 2011. Enabling multilevel trust in privacy preserving data mining. *TKDE* (2011).

[32] Yujia Li, Kevin Swersky, and Rich Zemel. 2015. Generative moment matching networks. In *ICML*.

[33] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *NIPS*.

[34] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In *NIPS*.

[35] Faisal Mahmood, Richard Chen, and Nicholas J Durr. 2018. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE transactions on medical imaging* (2018).

[36] V. K. Mansinghka, C. Kemp, J. B. Tenenbaum, and T. L. Griffiths. 2006. Structured Priors for Structure Learning. In *UAI*.

[37] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *ICCV*.

[38] Tom M Mitchell. 1980. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ. New Jersey.

[39] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. *ICLR* (2018).

[40] Klemen Naveršnik and Klemen Rojnik. 2012. Handling input correlations in pharmacoeconomic models. *Value in Health* (2012).

[41] P. Odom, T. Khot, R. Porter, and S. Natarajan. 2015. Knowledge-Based Probabilistic Logic Learning. In *AAAI*.

[42] Phillip Odom and Sriraam Natarajan. 2015. Active advice seeking for inverse reinforcement learning. In *AAAI*.

[43] Phillip Odom and Sriraam Natarajan. 2018. Human-guided learning for probabilistic logic models. *Frontiers in Robotics and AI* (2018).

[44] Michela Paganini, Luke de Oliveira, and Benjamin Nachman. 2018. CaloGAN: Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. *Physical Review D* (2018).

[45] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR* (2016).

[46] Ernest M Scheuer and David S Stoller. 1962. On the generation of normal random vectors. *Technometrics* (1962).

[47] Bernhard Schölkopf, Patrice Simard, Alex J Smola, and Vladimir Vapnik. 1998. Prior knowledge in support vector kernels. In *Advances in neural information processing systems*. 640–646.

[48] Rittika Shamsuddin, Barbara M Maweu, Ming Li, and Balakrishnan Prabhakaran. 2018. Virtual patient model: an approach for generating synthetic healthcare time series data. In *ICHI*.

[49] Jude W Shavlik and Geoffrey G Towell. 1989. Combining explanation-based learning and artificial neural networks. In *Proceedings of the sixth international workshop on Machine learning*. Elsevier.

[50] Geoffrey G Towell and Jude W Shavlik. 1994. Knowledge-based artificial neural networks. *Artificial intelligence* (1994).

[51] Yan Wang, Biting Yu, Lei Wang, Chen Zu, David S Lalush, Weili Lin, Xi Wu, Jiliu Zhou, Dinggang Shen, and Luping Zhou. 2018. 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. *NeuroImage* (2018).

[52] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. 2018. Face aging with identity-preserved conditional generative adversarial networks. In *CVPR*.

[53] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*.

[54] S. Yang and S. Natarajan. 2013. Knowledge Intensive Learning: Combining Qualitative Constraints with Causal Independence for Parameter Learning in Probabilistic Models. In *ECMLPKDD*.

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.