PRESERVATION OF ANOMALOUS SUBGROUPS ON VARIATIONAL AUTOENCODER TRANSFORMED DATA

Samuel C. Maina[1] *Reginald E. Bryant*[1] *William O. Ogallo*[1]
Kush R. Varshney[1] *Skyler Speakman*[1] *Celia Cintas*[1]
Aisha Walcott-Bryant [1] *Robert-Florian Samoilescu*[1,2] *Komminist Weldemariam*[1]

[1] IBM Research, Nairobi, Kenya
[2]Politehnica University of Bucharest, Bucharest, Romania

ABSTRACT

We investigate the effect of variational autoencoder (VAE) based data anonymization and its ability to preserve anomalous subgroup properties. We present a Utility Guaranteed Deep Privacy (UGDP) system which casts existing anomalous pattern detection methods as a new utility measure for data synthesis. UGDP's approach shows that properties of an anomalous subset of records, identified in the original data set, are preserved through the anonymization of a VAE. This is despite the newly generated records being completely synthetic. More specifically, the Bias-Scan algorithm identifies a subgroup of records that are consistently over- (or under-) risked by a black-box classifier as an area of 'poor fit'. This scanning process is applied on both pre- and post- VAE synthesized data. The areas of poor fit (i.e. anomalous records) persist in both settings. We evaluate our approach using publicly available datasets from the financial industry. Our evaluation confirmed that the approach is able to produce synthetic datasets that preserved a high level of subgroup differentiation as identified initially in the original dataset. Such a distinction was maintained while having distinctly different records between the synthetic and original dataset.

Index Terms— privacy, bias, anonymization, variational autoencoder

1. INTRODUCTION

Data-driven processes of identifying interesting and useful subgroups remains an important task across many industries. For example, financial services providers are increasingly relying on machine learning-based approaches for subgroup discovery in shaping investment and marketing strategies [1, 2], designing bespoke portfolio and insurance products [3, 4], managing risk [5], detecting fraud [6], and complying with anti-discrimination or fairness regulations [7]. In healthcare, attempts have been made in identifying and classifying patients' subgroups with similar prognoses and manifestations, and to understand responses to different treatment regimes in an attempt to personalize medical service provision [8, 9, 10, 11]. A key objective in these tasks is the discovery of client (patients) segments, or subgroups in individual-level data, defined by demographic, psychographic, behavioral, or other variables, that are of interest or anomalous according to some criterion [12].

Privacy is a critical consideration while working with individual-level data in different regions of the world and is protected by laws such as the GLBA [13], HIPAA [14], and FERPA [15] in the US and the GDPR [16] in Europe. Failure to appropriately protect patient or customer data exposes organizations to significant reputational and legal risks. Anonymization seeks to protect private and sensitive information of a client and or their activities. When sharing data with privileged partners, stakeholders or regulators, most institutions use common anonymization techniques (e.g., removal, redaction, encryption, data masking, perturbation, and generalization) to safeguard the privacy and secure the sanctity of the data.

Advances in generative machine learning such as variational autoencoders (VAEs) and generative adversarial networks (GANs) are being applied to the anonymization problem [17, 18, 19, 20, 21, 22, 23]. Their main idea is to learn the salient characteristics of the data distribution and sample new (synthetic) individuals from the distribution. The synthetic data from the generative models retain the properties of the original data and can, therefore, be used as a proxy and be shared without risks of re-identification or information leakage [19].

Different anonymization techniques distort the dataset in various ways and may not be desired depending on the downstream task to be performed on the synthesized data set [24]; as anonymization without preserving the data utility, i.e., the information content of the data is not desired [25]. In this paper, we examine the extent to which VAE anonymization techniques preserve subgroups of interest on a dataset. In particular, we investigate the persistence of bias in binary classifiers pre- and post synthesization. Therefore, our utility or figure of merit for an anonymization technique is its ability to yield the same or similar *most-anomalous* subgroup.

The subgroups of interest in this work are identified using Bias-Scan [26]. Bias-Scan identifies a subset of self-similar records that are consistently over (or under) risked by a black-box classifier. These records represent an area of 'poor fit' by the classifier. We apply Bias-Scan on the original data and the synthesized data generated by a VAE and investigate the proportion of the subgroup that was preserved.

We use the *Bank of Portugal* dataset [27] and the *Adult Dataset* [28] to evaluate our proposed approach. Our results indicate two things of note: First, from the perspective of privacy, the data synthesis procedure does not preserve identities of the most anomalous individual records. We, therefore, conclude that the VAE performs quite well in anonymizing the original dataset with minimal re-identification risk. Second, from the perspective of utility, the transformation process yields mid-value Jaccard distance values (about 0.5) for attribute-value overlap indicating that the overall statistical properties of the data are relatively preserved. With this, we are able to demonstrate our approach achieves anonymization on individual records and allows data reuse and sharing amongst stakeholders.

2. FRAMEWORK

In this section, we briefly describe the theoretical and algorithmic frameworks that we use for this work.

2.1. Variational Autoencoders

Autoencoders are generative models designed to capture the underlying distribution of input data and reproducing it using its essential underlying features. These essential features are lower-order representations of the data determined within the internal structure of the autoencoder. This lower-order or compressed representation is termed the *latent space*. As information about the original data is compressed, the output of autoencoders is not fundamentally the same as the input —the output only captures certain aspects of the original data. VAEs take this notion of inexact replication of the original data to another level. They are designed to produce variations of the input data, generating new synthetic data based on the underline statistics of the input data. Typically, used for image data reproduction, we are focused on using VAEs to reproduce tabular data as a way to represent the underlying statistics of the input data, thus establishing a data protection mechanism to ensure privacy.

2.2. Bias-Scan Algorithm

Bias-Scan takes the *subset scanning* approach to detecting bias in binary classifiers [26, 29] and treats the task as a search problem with the goal of finding the sub-population that is the most systematically over- (or under-) risked by a provided black-box classifier. Bias-Scan defines bias as the divergence between a model's predicted risks and the observed outcomes for a subset of records S. This definition is quantified as a log-likelihood ratio statistic based on the Bernoulli distribution. The null hypothesis is that all records in a subset have their binary outcome correctly captured by the classifier. The alternative hypothesis is that some subset of records have their binary outcomes generated by a multiplicative increase (or decrease) in their predicted odds $q\frac{p_i}{1-p_i}$. The scoring function maximized over all subpopoulations, S, in Bias-Scan is:

$$score_{bias}(S) = \max_{q} log(q) \sum_{i \in S} y_i - \sum_{i \in S} log(1 - \hat{p_i} + q\hat{p_i}),$$

where $\hat{p_i}$ is the estimated probability from the classifier and $q > 1$ is a multiplicative factor. Intuitively, any subset of records with a large number of $y_i = 1$ labels while also having a consistently small p_i predictions from the model will have a high score due to the divergence between the model's predictions and the observed outcomes for that particular subset. More formally, this subset shows the *most evidence* of being affected by the alternative hypothesis defined above. This scoring function satisfies the Linear Time Subset Scanning property (LTSS) [29] which allows for efficient maximization over relevant subsets of data. The LTSS property efficiently rules out provably-suboptimal subsets from the search process, drastically decreasing the relevant subsets under consideration from $O(2^n)$ to $O(n)$.

Bias-Scan does this efficient search over a multi-dimensional tensor with each feature (column in a dataset) being a mode in the tensor and each record (row in a dataset) falling into one of the *cells*. Bias-Scan identifies an *axis-aligned* subset of these cells such that the records in the cells maximize the scoring function. An example of axis-aligned subsets that span three modes of a tensor would be Race (Black, Hispanic), Gender (Female), Income (Low, Middle), i.e Black or Hispanic females who have low or middle income. Bias-Scan iteratively optimizes over each mode of the tensor until convergence to a local maximum is found. Exploiting the LTSS property of the scoring function guarantees that each optimization step over a mode in the tensor is done efficiently and exactly. For example; if we had a feature with 6 attribute values, an exhaustive search over combinations of values of this mode would require 2^6 scoring operations. However, this maximization can be done by only scoring 6 combinations while still guaranteeing that the optimal subset of values for that mode will be found. The joint optimization over all modes depends on the order in which the modes are optimized, and therefore multiple restarts are used to help explore the space and reach a global maximum.

In this work, we apply the Bias-Scan algorithm to identify the subgroup of interest, the most anomalous subgroup from the original (categorical or binned) data. We will then verify the extent to which this subgroup is preserved within the VAE generated synthetic dataset by running the Bias-Scan algorithm on this new dataset and comparing the derived anomalous subgroup with that from the original dataset.

3. UTILITY GUARANTEED DEEP PRIVACY (UGDP) SYSTEM

We present the Utility Guaranteed Deep Privacy (UGDP) system that identifies the most anomalous subgroup in a dataset by maximizing the bias between the groups predicted odds ratio from the model and the observed odds ratio from the data and preserve anomalous subgroups when transformed using a variational autoencoder. Figure 1 illustrates an overview of the UGDP system.

First, we trained two classifiers using the original datasets — a Random Forest predictive classifier and an XGBoost predictive classifier. Each classifier's parameters were optimized by cross-validation, with model performance measured using Area Under the ROC Curve (AUC). The classifiers were calibrated to improve their probability estimation and used to generate the predicted probabilities of a respondent taking up the proposed term deposit product. Second, we apply the Bias-Scan algorithm to identify the single subgroup of attribute space with the highest bias score. This is the subgroup of data that is most anomalous with regard to the predicted and the observed outcome. The bias score, therefore, reflects the extent to which the classifier makes the prediction error for the observations in this subgroup.

Third, following [30], we train a standard categorical variational autoencoder (VAE) to generate new samples of the original dataset. In our case, we use Adaptive Moment Estimation (ADAM) [31] as the optimization method, which computes adaptive learning rates for each parameter with a $LR = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. Here, the loss function is the reconstruction loss added to the K-L divergence. The input shape of the vectors varies depending on the dataset where all variables were encoded using one-hot encoding. We ran several simulations of the VAE on the original data to obtain 6 different sets of synthetic data samples. For each of the 6 synthetic data samples, we trained a Random Forest and an XGBoost model that was then used to obtain the respective predictive probabilities for each new dataset using the binary classifiers. Fourth, we applied the Bias-Scan algorithm to each of the 6 synthetic dataset samples to identify the most anomalous subgroup (subgroup with the highest bias score).

Lastly, for each classifier, we evaluated the pairwise dissimilarity between the individual records belonging to the most anomalous subgroup in the original dataset and the individual records belonging to the most anomalous subgroup in the 6 VAE synthetic samples. We

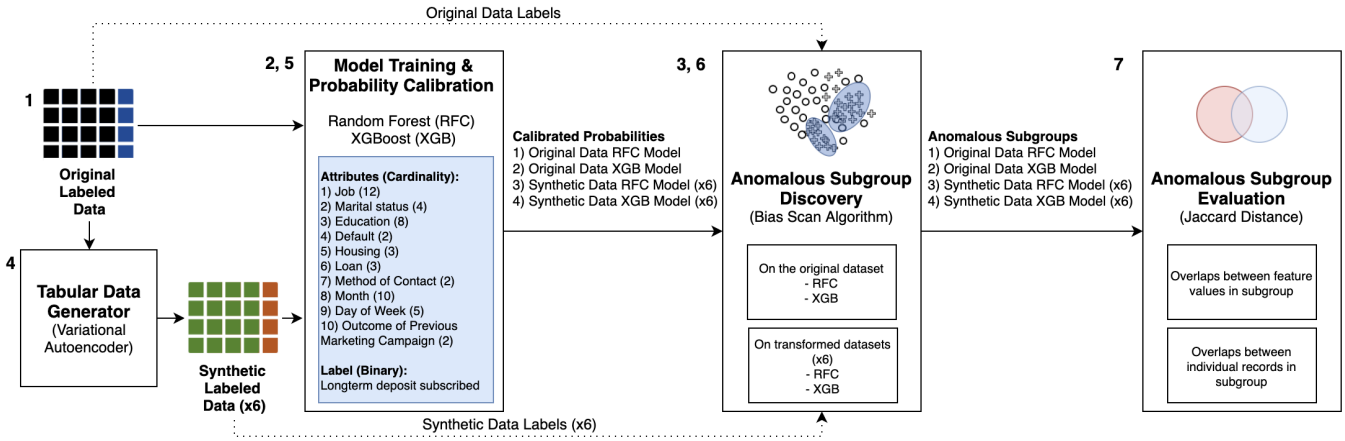

Fig. 1. Overview of the Utility Guaranteed Deep Privacy (UGDP) System. Numbers indicate steps as follows: (1) obtaining and pre-processing data, (2) training and calibrating predictive models on the original data, (3) employing bias-scan algorithm to discover subgroups on the original data, (4) generating tabular data using VAE (transformed data), (5) training and calibrating predictive models on the VAE-transformed data, (6) employing bias-scan algorithm to discover anomalous subgroup on the VAE-transformed data, and (7) evaluating the preservation of anomalous subgroup.

also evaluated the pairwise dissimilarity between the most anomalous subset of attribute values in the original dataset and the most anomalous subset of attribute values in the 6 VAE synthetic samples for each classifier. We quantified pairwise dissimilarity between sets of subgroups using the Jaccard distance, $d_{X,Y}$. The Jaccard distance is complementary to the Jaccard index, and is obtained by and is defined as: $d_{X,Y} = 1 - \frac{X \cap Y}{X \cup Y}$, where X and Y are sets of discrete elements and $0 \leq d_{X,Y} \leq 1$ with a higher values implying greater dissimilarity. We chose this dissimilarity measure since, by definition, we intended to compare finite discrete sets.

4. EXPERIMENT AND EMPIRICAL EVALUATION

For our experimental evaluation of our approach, we first use the publicly available data from a labeled direct marketing campaign of a Portuguese banking institution consisting of 10 customer attributes (a portion of the original 150) with the binary outcome of the acceptance of an offered long-term bank deposit product. This data was collected from May 2008 to November 2010 and consists of 41,188 records with 20 labeled attributes. This data describes whether a customer will accept a long-term bank deposit account (binary target label) [27]. Our preliminary experiment with this dataset takes a subset of the categorical attributes with both nominal and ordinal values. Specifically, we performed experimental evaluation on a filtered dataset that included the following 10 discrete attributes (listed with and their cardinality): *Job* (12); *Marital status* (4); *Education* (8); *Default* (2); *Housing* (3); *Loan* (3); *Method of Contact* (2); *Month* (10); *Day of Week* (5); and *Outcome of Previous Marketing Campaign* (2). The final dataset consists of $37,069$ records, 10 attributes, and 1 binary target label.

As shown in Table 1, we observe that the two predictive classifiers used in this study have relatively high AUC scores both the original data and the 6 synthetic samples. For example, Figure 2 illustrates the ROC curves for both for the original dataset and a sample VAE generated dataset (synthetic_5) under both classifiers and suggests that the performances of the random forest and XGBoost models for each dataset were comparable.

Model	Random Forest	XGBoost
Original	0.85	0.83
synthetic_0	0.779	0.789
synthetic_1	0.951	0.952
synthetic_2	0.946	0.949
synthetic_3	0.915	0.914
synthetic_4	0.900	0.903
synthetic_5	0.970	0.969

Table 1. AUC scores for the Random Forest and XGBoost predictive models on the original Bank of Portugal dataset and six synthetic VAE generated datasets.

The results of our investigation of the overlaps between the anomalous subgroups of individual records and attribute values in the original data and the VAE synthetic datasets are shown in Table 2. These results show the impact of the VAE transformation on the subgroup predictive bias of the classifiers. The Jaccard distances for individual records indicate little to no overlap between individual records of the subgroup most affected by the predictive bias of Random Forest and XGBoost classifiers. The Jaccard distances for the attribute values suggest a considerable overlap between subsets of the attribute space that are most affected by predictive bias.

This observation is exemplified in Figure 3 which compares the overlap between the most anomalous subgroups in the original dataset and a sample VAE generated synthetic dataset (synthetic_5) under the random forest classifier. As shown in Figure 3(a), there is no overlap (Jaccard distance = 1) between the individual records of the subgroup most affected by the predictive bias of the classifier as measured by the Jaccard distance metric which shows values very close to one. This suggests that the VAE performs relatively well in data anonymization and re-identification of individuals is not guaranteed. However, we observe that for the attribute (feature) values, there is significant overlap (Jaccard distance = 0.462) between the most anomalous subgroups of the model as shown in Figure 3(b). This implies that the characteristics of subgroups of interest in a

Category	Individual Records		Attribute Values	
model	RF	XGB	RF	XGB
synthetic_0	0.965	0.953	0.571	0.556
synthetic_1	0.949	0.958	0.481	0.536
synthetic_2	0.959	0.958	0.516	0.543
synthetic_3	0.949	0.975	0.469	0.545
synthetic_4	0.947	0.951	0.438	0.441
synthetic_5	1.000	1.000	0.462	0.581

Table 2. Dissimilarity, quantified by the Jaccard distance, between subgroups in the original dataset vs 6 VAE synthetic samples that are most affected by bias under different setups. Higher values imply greater dissimilarity

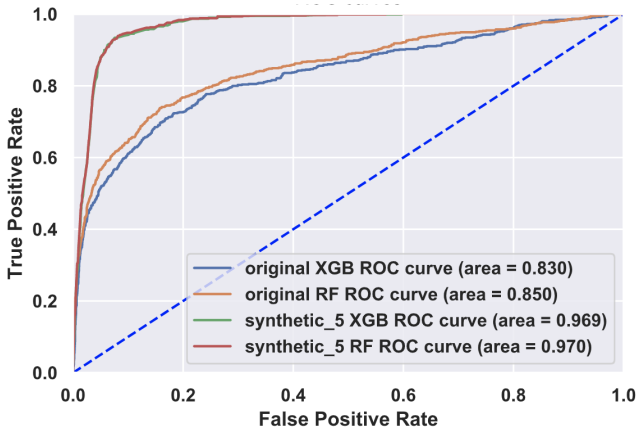

Fig. 2. Receiver Operating Characteristic (ROC) curves for Random Forest (RF) and XGBoost (XGB) classifiers trained on the original Bank of Portugal dataset and a sample VAE synthetic dataset.

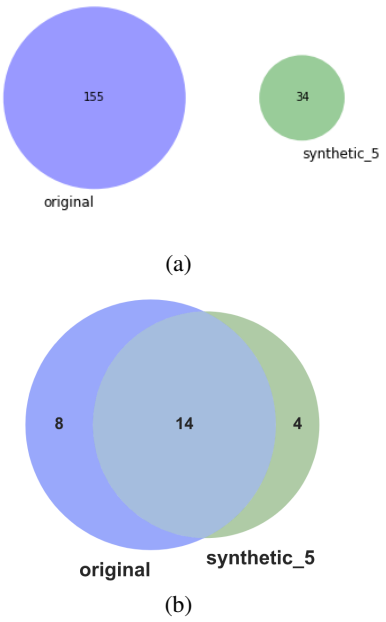

Fig. 3. (a) shows for a Random Forest classifier trained on the original Bank of Portugal data and on a synthetic data generated by the VAE, with no overlap between individual records of the subgroup most affected by the predictive bias of the classifier. (b) shows for the same setup, there is considerable overlap (Jaccard distance = 0.462).

dataset would not be lost when the data is transformed using VAEs.

Finally, we further tested the UGDP system using our second dataset, namely the Adult Dataset [28] (the 1994 (United States) Census-bureau dataset), to validate the approach with respect to preserving subgroup properties and privacy on transformed data through synthesis techniques. This dataset consists of 7 attributes from the original 14 features that are used to predict whether a person's income is higher or lower than $50k per year threshold. In particular, we observed minimal overlap between the individual records of the subgroups most affected by the predictive bias of the Random Forest (Jaccard distance = 0.96) and the XGB model (Jaccard distance = 0.89). Similar to the results observed in the Bank of Portugal data experiment, there was more overlap in the attribute values of the most anomalous subgroup for the Random Forest model (Jaccard distance = 0.65) and the XGB model (Jaccard distance = 0.43).

5. DISCUSSION

In this paper, we have presented the standard VAE as a tool for anonymization and utility preservation. In particular, we have demonstrated that this model achieves anonymization on individual records which allows for data reuse and sharing amongst the key stakeholders and clients. This is confirmed by the mid-value Jaccard distance on the overlaps for the anomalous subsets of individual records for the original and synthetic datasets after the data has been run through a classifier and the Bias-Scan algorithm. We have also shown that the VAE synthesizer preserves *some global* statistical distributional properties of the data as demonstrated by the mid-level values of the Jaccard distance metric for the feature values. We, therefore, conclude that the data transformation and synthesization preserves (to an extent) some key subgroup properties of the data.

A potential direction for further extensions of this work could be on developing a disciplined approach that can attain higher subgroup overlaps of attribute values between the two data sets while still achieving the high privacy levels as represented by minimal overlap at the level of the individual records. An initial approach would be modifying the standard VAE for data generation to a case where we incorporate constraints on how each attribute can vary with respect to one another when generating the synthetic data from the VAE. We are also exploring alternative generative approaches such as GANs [19, 32] to synthesize tabular data and check for consistency values between different attributes while preserving the privacy of individual records.

6. REFERENCES

[1] Marcos Lopez de Prado and Michael J. Lewis, "Detection of false investment strategies using unsupervised learning methods," *Quantitative Finance*, vol. 19, no. 9, pp. 1555–1565, 2019.

[2] Ning Sun, Jacqueline G. Morris, Jian Xu, Xiufang Zhu, and Ming Xie, "iCARE: A framework for big data-based banking customer analytics," *IBM Journal of Research and Development*, vol. 58, no. 5/6, pp. 4, Sept./Nov. 2014.

[3] Chris Lamberton, Damiano Brigo, , and Dave Hoy, "Impact

of robotics, RPA and AI on the insurance industry: Challenges and opportunities," *Journal of Financial Perspectives*, vol. 4, no. 1, pp. 8–20, May 2017.

[4] David Thesmar, David Sraer, Lisa Pinheiro, Nick Dadson, Razvan Veliche, and Paul Greenberg, "Combining the power of artificial intelligence with the richness of healthcare claims data: Opportunities and challenges," *PharmacoEconomics*, vol. 37, no. 6, Jun 2019.

[5] Bart van Liebergen, "Machine learning: A revolution in risk management and compliance?," *Journal of Financial Transformation*, vol. 45, pp. 60–67, 2017.

[6] Anuj Sharma and Prabin Kumar Panigrahi, "A review of financial accounting fraud detection based on data mining techniques," *International Journal of Computer Applications*, vol. 39, no. 1, pp. 37–47, 2012.

[7] Matthew Adam Bruckner, "The promise and perils of algorithmic lenders' use of big data," *Chicago-Kent Law Review*, vol. 93, no. 1, 2018.

[8] Hung-Chia Chen, Wen Zou, Tzu-Pin Lu, and James J. Chen, "A composite model for subgroup identification and prediction via bicluster analysis," *PLoS One*, vol. 9, no. 10 e111318, 2014.

[9] Rachel Knevel and Tom WJ Huizinga, "On using machine learning algorithms to define clinically meaningful patient subgroups," *Annals of the Rheumatic Diseases*, 2019.

[10] Joo Pedro Ferreira, Kevin Duarte, John J.V. McMurray, Bertram Pitt, Dirk J. van Veldhuisen, John Vincent, Tariq Ahmad, Jasper Tromp, Patrick Rossignol, and Faiez Zannad, "Data-driven approach to identify subgroups of heart failure with reduced ejection fraction patients with different prognoses and aldosterone antagonist response patterns," *Circulation: Heart Failure*, vol. 11, no. 7, 2018.

[11] Anne Molgaard Nielsen, Peter Kent, Lise Hestbaek, Werner Vach, and Alice Kongsted, "Identifying subgroups of patients using latent class analysis: should we use a single-stage or a two-stage approach? a methodological study using a cohort of patients with low back pain," *BMC Musculoskeletal Disorder*, vol. 18, no. 57, 2017.

[12] Stefan Rueping, "Ranking interesting subgroups," in *Proceedings of the International Conference on Machine Learning*, June 2009, pp. 913–920.

[13] "Gramm-Leach-Bliley Act (GLBA)," `shorturl.at/rIOV4`.

[14] "Health Insurance Portability and Accountability Act (HIPAA)," `shorturl.at/cqxy7`.

[15] "The Family Educational Rights and Privacy Act (FERPA)," `https://www.law.cornell.edu/uscode/text/20/1232g`.

[16] "General Data Protection Regulation (GDPR)," `https://eugdpr.org`.

[17] Yujia Li Max Welling Richard Zemel Christos Louizos, Kevin Swersky, "The variational fair autoencoder," in *Proceedings of the International Conference on Learning Representations*, 2016.

[18] Harrison Edwards and Amos Storkey, "Censoring representations with an adversary," in *Proceedings of the International Conference on Learning Representations*, 2016.

[19] 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*, vol. 11, no. 10, pp. 1071–1083, 2018.

[20] R. Bryant, C. Cintas, I. Wambugu, A. Kinai, A. Diriye, and K. Weldemariam, "Evaluation of Bias in Sensitive Personal Information Used to Train Financial Models," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Nov 2019, IEEE.

[21] Ho Bae, Dahuin Jung, and Sungroh Yoon, "AnomiGAN: Generative adversarial networks for anonymizing private medical data," 2020.

[22] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth, "DeepPrivacy: A generative adversarial network for face anonymization," 2019.

[23] Samuel Ainsworth, Nicholas J. Foti, Adrian K. C. Lee, and Emily B. Fox, "Interpretable VAEs for nonlinear group factor analysis," 2018.

[24] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney, "Distribution-preserving k-anonymity," *Statistical Analysis and Data Mining*, vol. 11, no. 6, pp. 253–270, 2018.

[25] LING LIU and M. TAMER ÖZSU, Eds., *Information*, pp. 1476–1476, Springer US, Boston, MA, 2009.

[26] Zhe Zhang and David Neill, "Identifying significant predictive bias in classifiers," *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[27] Sérgio Moro, Paulo Cortez, and Paulo Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, 2014.

[28] "US Adult Census Dataset," `https://archive.ics.uci.edu/ml/datasets/adult`.

[29] Daniel B. Neill, "Fast subset scan for spatial pattern detection," *Journal of the Royal Statistical Society Series B*, vol. 74, no. 2, pp. 337–360, 2012.

[30] Eric Jang, Shixiang Gu, and Ben Poole, "Categorical reparameterization with Gumbel-softmax," 2017.

[31] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2014.

[32] Ashutosh Kumar, Arijit Biswas, and Subhajit Sanyal, "eCommerceGAN: A generative adversarial network for e-commerce," 2018.