Topology for preserving feature correlation in tabular synthetic data.

ARIFEEN, M. and PETROVSKI, A.

2022

Topology for Preserving Feature Correlation in Tabular Synthetic Data

1st Murshedul Arifeen
School of Computing
Robert Gordon University
Aberdeen, Scotland
d.arifeen@rgu.ac.uk

2nd Andrei Petrovski
School of Computing
Robert Gordon University
Aberdeen, Scotland
a.petrovski@rgu.ac.uk

***Abstract*—Tabular synthetic data generating models based on Generative Adversarial Network (GAN) show significant contributions to enhancing the performance of deep learning models by providing a sufficient a mount o f t raining data. However, the existing GAN-based models cannot preserve the feature correlations in synthetic data during the data synthesis process. Therefore, the synthetic data become unrealistic and creates a problem for certain applications like correlation-based feature weighting. In this short theoretical paper, we showed a promising approach based on the topology of datasets to preserve correlation in synthetic data. We formulated our hypothesis for preserving correlation in synthetic data and used persistent homology to show that the topological spaces of the original and synthetic data have dissimilarity in topological features, especially in 0^{th} and 1^{st} Homology groups. Finally, we concluded that minimizing the difference in topological features can make the synthetic data space locally homeomorphic to the original data space, and the synthetic data may preserve the feature correlation under homeomorphism conditions.**

***Index Terms*—Synthetic Data, Correlation, GAN, Topology, Persistent Homology**

I. Introduction

Data-driven algorithms like deep learning models require a large dataset during training for learning the underlying pattern and enhancing performance [1]. However, many practical applications (water treatment plant, smart farming etc.) involve difficulties a nd t ime c onsumption f or g athering t he required training data, which creates sampling noise and difficulties in model training. An alternative solution is the generation of synthetic data from a small real-world dataset. Synthetic data provides efficient m odel t raining, e liminates sampling noise and bias, highly scalable, and reduces time consumption during data collection [1].

Synthetic data are of various types, including images, signals, or tabular, while this paper only considers tabular synthetic data generation. A popular model available in the literature for generating tabular synthetic data is GAN. For instance, Jordon et al. [2] proposed a high-quality differentially private synthetic data generation model based on GAN for Electronic Health Record (EHR) data. Choi et al [3] also proposed a GAN-based synthetic EHR data generation model. GAN is also used for synthesizing tabular data in [4] [5] [6] [7]. However, these data synthesis methods aim to anonymize health care data to preserve users' privacy. On the contrary, Xu et al. [8] designed a GAN-based general-purpose tabular data synthesis model named CT-GAN (Conditional Tabular GAN) by introducing mode-specific normalization. The CT-GAN can perform very well in generating synthetic data by preserving the underlying data distribution. Though there is a significant improvement for GAN-based synthetic data generation models, these models can not properly preserve the correlation among the features in synthetic data. Few papers [7] [5] consider correlation preservation with data anonymization, yet high error exists between original and synthetic data correlation matrix. Therefore, these models are suitable where data privacy is the first priority but may fail where correlation is necessary.

Feature correlation shows significant importance in various data analysis and modeling tasks like correlation-based feature weighting is widely applied for soft sensor modeling [9] [10] where synthetic data is highly required. Moreover, few feature selection strategies highly depend on correlation coefficient values, and most importantly, synthetic data should be as realistic as original data by preserving all the inherent characteristics (feature correlation, manifold, temporal correlation). Therefore, GAN-based synthetic data generation by preserving all the original characteristics of the real dataset can further enhance the synthetic data generation works.

In this paper we particularly investigate the following research question for feature correlation preservation in synthetic data and presented preliminary results of our experiments.

Research question: *Can we preserve pairwise feature correlation of the synthetic data by making the underlying topological space of the synthetic data locally homeomorphic to the underlying topological space of the original data?*

In section II, a brief review work is presented. Then in section III, we discussed the reasoning of our hypothesis with theoretical analysis. Section IV shows the methodology for constructing the topological space for original and synthetic data. In section V we demonstrated the experiments and discussed the results that support our hypothesis. Finally, section VI concludes this work in progress paper with future research plans.

TABLE I: Overview of the most relevant works for tabular data synthesis

Ref	Year	Model	Aim	Contribution	Comments
[3]	2017	GAN	Privacy preservation	Minibatch averaging algorithm to overcome GAN over-fitting.	Quantitative analysis is missing for evaluating pairwise feature correlation.
[6]	2018	DCGAN	Privacy preservation	Information loss and classification loss is combined to realize connection between features and labels.	Correlation between labels and other attributes are considered but correlation among feature pairs are not considered.
[7]	2018	GAN	General data synthesis	Correlation preservation, multimodal data learning.	High error between original and synthetic pairwise feature correlation in synthetic data.
[2]	2019	GAN	Privacy preservation	Differential privacy based synthetic data generation with a new evaluation criteria.	Explicit evaluation of feature correlation preservation is missing.
[8]	2019	CTGAN	General data synthesis	Mode specific normalization for generating multi modal synthetic data	Fails to retain the pairwise feature correlations.
[5]	2020	CGAN	Privacy preservation	Convolutional neural network for capturing inter feature correlation.	Quantitative analysis is missing for evaluating pairwise feature correlation.
[4]	2021	CTGAN	Privacy preservation	Pearson for numerical correlation measurement.	Pearson fails to capture nonlinear relationship among the features.

Legend:Ref–References, CTGAN– Conditional Tabular GAN, CGAN–Convolutional GAN, DCGAN–Deep Convolutional GAN.

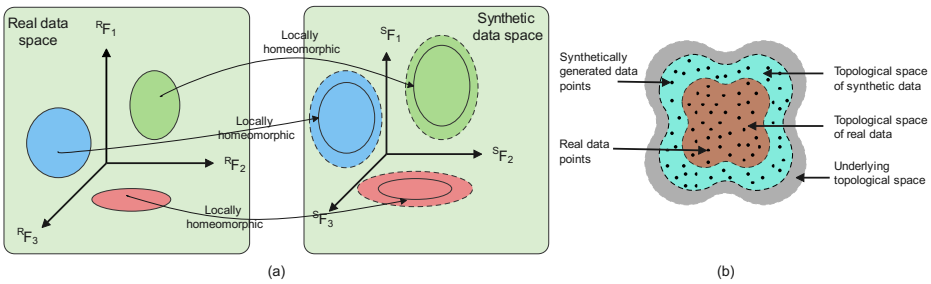

Fig. 1: Figure (a) illustrates the concept of making the synthetic data space locally homeomorphic to the original data space. To visualize, here we considered 3 dimensional space with 3 features. Figure (b) represents that the topological space of real data is a subset of the topological space of the synthetic data and both synthetic and real data space is the subset of the underlying topological space.

II. Literature Review

GAN was first introduced by Ian Goodfellow [11]. The GAN model comprises two modules known as generator and discriminator. The generator generates possible data instances from noise values, and these data instances become negative samples for the discriminator. On the contrary, the discriminator learns to differentiate the synthetic data samples of the generators from the original data. Also, the generator gets feedback from the discriminator for generating synthetic data. Throughout the training process, the output of the generator becomes realistic, and the discriminator can no longer differentiate between the generated samples and the original samples. At this stage, the training stops. In recent years, GAN has been favored for generating synthetic data in various domains like computer vision. GAN also gets popularity for tabular data generation to increase the number of data samples. Table 1 depicts an overview of the existing GAN-based synthetic data generating models. We can see the models aim to preserve privacy or data anonymization, but these models ignore preserving the original data characteristics. The synthetic data without the characteristics of the original data becomes unrealistic. Therefore, the original data characteristics preservation in synthetic data is required for various applications, including correlation-based feature selection, time series data analysis, and many more.

III. Hypothesis

The pairwise correlation among the features of original data could be preserved in the synthetic data by making the underlying topological space of the synthetic data locally homeomorphic to the underlying topological space of the original data.

A. Reasoning for the hypothesis

Correlation can be viewed as an intrinsic relationship among various dimensions (features) of the data points. By nature, a correlation exists among the features of the data points collected using electronic devices. For instance, the features pH value, water level and soil moisture of some data points show a high positive correlation, and this relationship among the features exist naturally. Therefore, preserving the correlation in synthetic data could be possible by preserving the shape of the original data. This last statement is justified below.

Let A represents the original dataset and $\mathcal{A}$ be the underlying topological space of A. Also, consider B defines the generated synthetic data, and the underlying topological space of B is $\mathcal{B}$. If it is possible to make space $\mathcal{B}$ homeomorphic to space $\mathcal{A}$, then the original data samples A and synthetic data samples B should demonstrate identical characteristics. Since space $\mathcal{B}$ is a continuous deformation of space $\mathcal{A}$ under the homeomorphism condition, the nearness relationship of

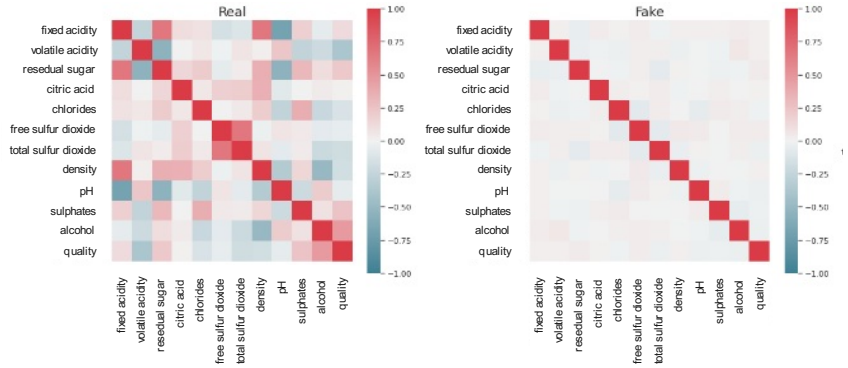

Fig. 2: Left matrix shows pairwise feature correlation of the original WQ dataset, while the right matrix shows pairwise correlation of the synthetic WQ dataset.

the data points is preserved in both spaces. Mathematically, we can think the real dataset A is the subset of the synthetic dataset B and both A and B are the subsets of the underlying spaces $\mathcal{A}$ or $\mathcal{B}$, i.e $A \subset B \subset \mathcal{A}/\mathcal{B}$. This idea is illustrated in figure 1. Moreover, correlation between pair of features is related to the nearness relationships of the data points in the space. Hence, preserving the shape (topological space) will preserve the nearness among data points which should further retain the correlation behavior.

B. Correlation and nearness

Lets consider a dataset with two features X and Y. Then, the correlation between X and Y with n samples can be represented as, $r(X,Y) = \frac{\frac{1}{n}\sum_i x_i y_i - \mu_x \mu_y}{\sigma_x \sigma_y}$ where, μ_x and μ_y denotes the mean of X and Y, σ_x and σ_y denotes standard deviations. If the dataset of features X and Y is standardized meaning $\mu = 0$ and $\sigma = 1$ then the correlation becomes, $r(X^*, Y^*) = \frac{1}{n}\sum_i x_i y_i$.

The nearness between n data points of features X and Y can be defined using a distance metric. Euclidean metric is considered here for showing the relationship between correlation and nearness as follows [12]-

$$d(X,Y) = \sqrt{\sum_i^n (x_i - y_i)^2}$$
$$= \sqrt{\sum_i^n x_i^2 + \sum_i^n y_i^2 - 2\sum_i^n x_i y_i}$$

If the data is standardized then $\sum_i^n x_i^2 = \sum_i^n y_i^2 = n$

$$d(X^*, Y^*) = \sqrt{n + n - 2\sum_i^n x_i y_i}$$
$$= 2n(1 - \frac{1}{n}\sum_i^n x_i y_i)$$
$$= 2n(1 - r(X^*, Y^*))$$
$$r(X^*, Y^*) = 1 - \frac{d^2(X^*, Y^*)}{2n}$$

From the above analysis, we can see that correlation between pair of features is inversely proportional to the distances among the data points. Hence, neighbor data points show similar correlation value for the features.

C. Homeomorphism

Under homeomorphism conditions, two spaces are topologically equivalent and it can be defined as follows [13]-

Definition A homeomorphism between topological spaces $\mathcal{A}$ and $\mathcal{B}$ is a bijective map $f : \mathcal{A} \to \mathcal{B}$ such that f and it's inverse f^{-1} are both continuous.

In other words, for all open sets $U \in \mathcal{A}$, $f(U)$ must be open in $\mathcal{B}$. Similarly, for all open sets $V \in \mathcal{B}$, $f^{-1}(V)$ must be open in $\mathcal{A}$. Since the open set U is neighbor of each of its points say $x \in \mathcal{A}$, U must contain an open set P such that $x \in P \subset U$. Moreover, f is continuous therefore $f(P)$ must be open in $\mathcal{B}$. From this we can conclude that $f(U)$ is a neighbor of $f(x)$ in $\mathcal{B}$, since $f(x) \in f(P) \subset f(U)$. In a similar way, we can show that $f^{-1}(V)$ is a neighborhood of x in $\mathcal{A}$. Therefore, under homeomorphism condition, both spaces preserved the neighborhood property of the data points.

IV. Methodology

A. Vietoris-Rips Complex

The tabular dataset is discrete, while the topological space is continuous. Therefore, a simplicial complex is used to build topological spaces from the discrete data points (point clouds). Simplicial complex can be constructed using the combination of different simplices.

1) Simplex, Simplices and Simplicial Complex: A k-simplex σ can be defined as a k-dimensional polytope which is a convex hull of its $k+1$ vertices, say v_0, v_1, .., $v_k \in \Re^d$[14]. The dimension of σ_k is $dim\sigma = k$. Simplices is just the plural form of a simplex. Simplices can be of different types, starting from $0-simplex$ (vertex), $1-simplex$ (edge), $2-simplex$ (triangle), $3-simplex$ (tetrahedron), and so on. On the contrary, a simplicial complex that represents a topological space can be constructed using multiple simplices. A simplicial complex κ can be defined as a set of simplices

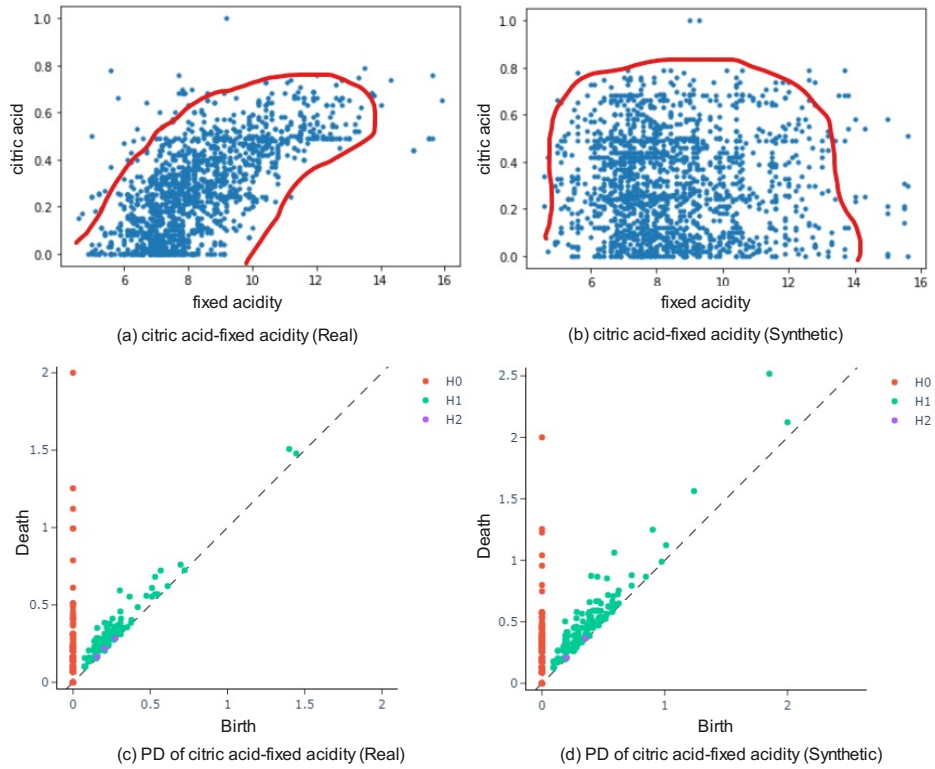

Fig. 3: This figure illustrates the scatter plot (first row) and PD (second row) of the citric acid-fixed acidity feature pair of WQ dataset for both real and synthetic data samples. A high positive correlation can be observed in figure (a) for real data samples but the correlation has been broken in the synthetic counterpart of the feature pair (figure (b)). Figure (c) and (d) depicts the topological features (0^{th}, 1^{st}, and 2^{nd} homology groups) of point clouds of figure (a) and (b) respectively.

that obeys two conditions- (1) Every face of $\sigma \in \kappa$ is also in κ and (2) if σ_i, $\sigma_j \in \kappa$, then $\sigma_i \cap \sigma_j$ is a face of both σ_i and σ_j, where a face of a simplex σ_k is the convex hull of any non-empty subset of points (v_0, v_1, .., v_k) [14].

A Vietoris-Rips Complex (VR) is a type of an abstract simplicial complex which is widely used in topological data analysis for constructing topological spaces from the point clouds. In this paper, we used the VR to construct the underlying topological space of the data points (both for original and synthetic data points). Let's say X represents the data points, then the VR of X denoted as $VR_d(X)$ contains the set of all simplices of X where the elements of X, $(x_1, x_2, ..x_k)$ satisfy the metric relation of $dist(x_i, x_j) < d$ for all i, j. Also, VR satisfy the nested relation of subcomplexes and it is possible to track the changes in topological features (connected components, loops and voids) as the metric d changes, through persistent diagram. Therefore, as an abstract simplicial complex, VR complex can be used to generate topological spaces from the discrete data points [15] [16].

B. Persistent Homology

Persistent homology [15] is the mathematical tool for understanding the topological features of a topological space constructed using simplicial complexes. It is an extension of the homology to the filtered chain complex settings [15]. The simplicial homology assigns homology groups to the simplicial complex for extracting topological features. For instance, the 0^{th} homology groups denote the topological features of connected components. Similarly, the 1^{st} and 2^{nd} homology groups represent the loops and voids topological features of the topological spaces respectively. Persistent homology reflects the homology groups through the Persistent diagrams (PD) of the simplicial complexes. Therefore, topological features are the heart of understanding topological space and PD is the tool used for visualizing the topological features. Moreover, we can use Wasserstein distance to compute the dissimilarity between two PD. Persistent homology has well established mathematical foundations and getting popular for data analysis by inspecting topological features. The essence of persistent homology is described here for understanding our hypothesis. A deep mathematical analysis of persistent homology and related topological data analysis concepts are presented in [15] [16].

V. Preliminary Experiments and Results

To support our hypothesis and understand the topological dissimilarity of the original and synthetic topological spaces of tabular data, we used python Giotto-TDA library and the following dataset.

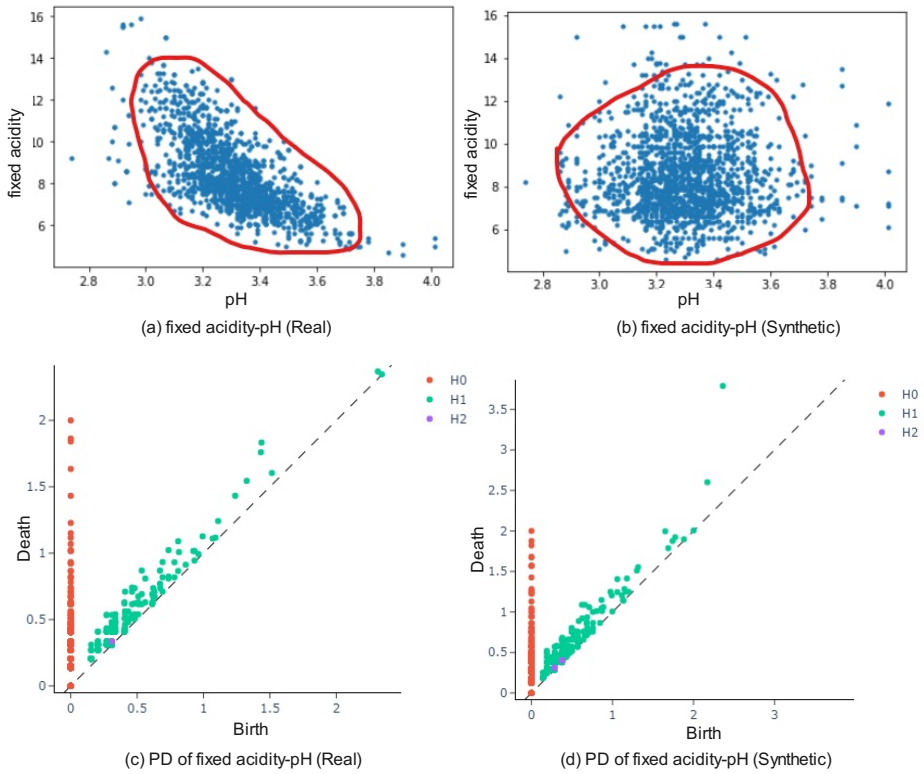

Fig. 4: This figure illustrates the scatter plot (first row) and PD (second row) of the fixed acidity-pH feature pair of WQ dataset for both real and synthetic data samples. A high negative correlation can be observed in figure (a) for real data samples but the correlation has been broken in the synthetic counterpart of the feature pair (figure (b)). Figure (c) and (d) depicts the topological features (0^{th}, 1^{st}, and 2^{nd} homology groups) of point clouds of figure (a) and (b) respectively.

TABLE II: Dissimilarity in topological features captured by the persistent diagrams of real and synthetic dataset

Dataset	**Feature pairs**	0^{th} **Homology**	1^{th} **Homology**	2^{nd} **Homology**
WQ	Citric acid- fixed acidity	1.96521339	0.76831807	0.01224039
	Citric acid- total sulfur dioxide	0.61243486	0.18172829	0.00710125
	fixed acidity-pH	2.01292377	1.03884327	0.03319647
Anomaly	Butane-Carbon Monoxide	4.84583870	1.52988194	0.00119914808
	Humidity-Butane	5.12984892	1.54729505	0.01191398
	Temperature- Humidity	2.04251361	0.483071038	0.00181607064
SRU	Gas flow- Air flow	3.83539682	1.41767874	0.01166989
	Secondary air flow- SWS air flow	1.74629258	0.442073456	0.000455793879
	Gas flow- secondary air flow	4.03835353	1.05204137	0.01010775

A. Dataset

For evaluating our hypothesis, we considered three industrial process dataset such as Wine Quality (WQ) [17], Anomaly [18], and Sulfur Recovery Unit (SRU) [19]. We chose three feature pairs from each dataset to illustrate the topological dissimilarity. Table II shows the dataset with the chosen feature pairs.

B. Synthetic Data Generation

First, we imported the CT-GAN model from the Github repository and generated synthetic data for each of the dataset. Then we examined the correlation matrix for each of the dataset and its synthetic counterparts. For example, figure 2 shows the correlation matrix for WQ dataset (both for real and synthetic). It is evident from this figure that the pairwise correlations among the features are not preserved in the synthetic data generated by the CT-GAN model. Moreover, the first rows of figure 3 and 4 show the scatter plots of the two feature pairs (citric acid-fixed acidity, fixed acidity-pH) of the WQ dataset and their synthetic counterparts. It is obvious from these figures that the neighborhood relationship among the data points is broken in the synthetic data, for instance, figure 3(a) shows a high positive correlation (increasing trends for both features) among the citric acid-fixed acidity feature pair, but this behavior is broken in the synthetic space (figure 3(b)). Similarly, figure 4(a) shows a high negative correlation (decreasing trends for both features) among the fixed acidity-pH feature pair, but this behavior is not preserved in the

synthetic space (figure 4(b)).

C. Persistent Diagram Computation

We first transformed the original and synthetic data into point clouds to construct corresponding VR complexes. Then the VR function is called from the Giotto-TDA library to generate the underlying topological spaces. The function returns two arrays- one array for the original dataset and the other one for the synthetic dataset. These arrays contain the birth-death value pairs for each topological feature generated throughout the VR complex construction process. After that, we used the PD plotting function for visualizing the topological features. The second rows of figure 3 and 4 illustrate the PD of the feature pairs of the first rows (both for real and synthetic data). From these PDs, we can see the 0^{th} Homology groups (connected components) and few components of the 1^{st} Homology groups (loops) are the significant topological features of the underlying topological spaces of the real and synthetic data. However, some components of 1^{th} and all the components of 2^{nd} Homology groups are generated along the diagonal and can be considered as noise. Finally, the Wasserstein distance is used to compute the difference between real and synthetic PD pairs for the corresponding feature pairs of the datasets. Table II shows the dissimilarity in topological features (Homology groups) for the feature pairs considered from each dataset.

VI. Discussion and Conclusion

Our hypothesis claims that making the underlying topological spaces of the original and the synthetic data locally homeomorphic may preserve pairwise feature correlation in synthetic space. In other words, there will be dissimilarity in the topological features between original and synthetic data generated by CT-GAN, since the CT-GAN can not preserve feature correlation. From our primary experiments, we found a significant difference in 0^{th} Homology groups and a small difference in 1^{st} Homology groups. However, the 2^{nd} Homology group can be discarded as the difference is negligible (table II). Therefore, minimizing the difference in the Homology groups may preserve the topological space of both original and synthetic data. Strictly speaking, minimizing topological differences will preserve the shape of the original data in synthetic space and under homeomorphism conditions, the synthetic space will preserve the nearness relationships among the data points. Since correlation is related to the neighborhood property of data points, synthetic data locally homeomorphic to the original data should preserve the correlation among the features.

In conclusion, preserving topological space can be a promising way to preserve the feature correlations in synthetic data. In our future work, we will create a topological loss function and use this loss function to fine-tune the CT-GAN model so that the GAN minimizes the difference in homology groups and learns the shape of the data. We are expecting that the training of CT-GAN through topological loss function will generate synthetic data with the same feature correlation as the original data. We will also analyze the complexity of the CTGAN model for generating tabular synthetic data under topological loss term.

VII. Acknowledgement

This work is a part of the IoT based Inferential Measurement for Asset Integrity and Security project, and supported by School of Computing, Robert Gordon University, Aberdeen, Scotland.

References

[1] "Synthetic data generation: 3 key techniques and tips for success," Jun 2022. [Online]. Available: https://datagen.tech/guides/synthetic-data/synthetic-data-generation/

[2] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in *International conference on learning representations*, 2018.

[3] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," in *Machine learning for healthcare conference*. PMLR, 2017, pp. 286–305.

[4] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, "Ctab-gan: Effective table data synthesizing," in *Asian Conference on Machine Learning*. PMLR, 2021, pp. 97–112.

[5] A. Torfi and E. A. Fox, "Corgan: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records," in *The Thirty-Third International Flairs Conference*, 2020.

[6] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *arXiv preprint arXiv:1806.03384*, 2018.

[7] L. Xu and K. Veeramachaneni, "Synthesizing tabular data using generative adversarial networks," *arXiv preprint arXiv:1811.11264*, 2018.

[8] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, "Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted sae," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3235–3243, 2018.

[10] X. Yuan, L. Feng, Y. Wang, and K. Wang, "Stacked attention-based autoencoder with feature fusion and its application for quality prediction," in *2021 IEEE 10th Data Driven Control and Learning Systems Conference (DDCLS)*. IEEE, 2021, pp. 1368–1373.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[12] S. Borgatti, "Measures of Similarity and Distance." [Online]. Available: http://www.analytictech.com/mb876/handouts/distance_and_correlation.htm

[13] W. A. Sutherland, *Introduction to metric and topological spaces*. Oxford University Press, 2009.

[14] J. Xu, X. Li, and H. Wang, "Extracting topological features from big data using persistent density entropy," *Journal of Physics: Conference Series*, vol. 1168, p. 032017, feb 2019. [Online]. Available: https://doi.org/10.1088/1742-6596/1168/3/032017

[15] F. Hensel, M. Moor, and B. Rieck, "A survey of topological machine learning methods," *Frontiers in Artificial Intelligence*, vol. 4, p. 681108, 2021.

[16] F. Chazal and B. Michel, "An introduction to topological data analysis: fundamental and practical aspects for data scientists," *Frontiers in artificial intelligence*, vol. 4, 2021.

[17] "Wine quality." [Online]. Available: https://www.kaggle.com/datasets/rajyellow46/wine-quality

[18] P. Jain, S. Jain, O. R. Zaïane, and A. Srivastava, "Anomaly detection in resource constrained environments with streaming data," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 3, pp. 649–659, 2021.

[19] L. Fortuna, S. Graziani, A. Rizzo, M. G. Xibilia *et al.*, *Soft sensors for monitoring and control of industrial processes*. Springer, 2007, vol. 22.