Human Systems Engineering and Design (IHSED2025), Vol. 198, 2025, 477–484
https://doi.org/10.54941/ahfe1006801

Data Synthetization and Feature Analysis: A Study in Bladder Cancer Recurrence Data

Sandi Baressi Šegota[1], Ivan Lorencin[2], Nikola Anđelić[1], Vedran Mrzljak[1], Antun Gršković[3,4], Juraj Ahel[3,4], Klara Smolić[3,4], and Dean Markić[3,4]

[1]Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia
[2]Faculty of Informatics, Juraj Dobrila University of Pula, Zagrebačka 30, 52100 Pula, Croatia
[3]Department of Urology, Clinical Hospital Center Rijeka, Krešimirova 42, 51000 Rijeka, Croatia
[4]Faculty of Medicine, University Of Rijeka, Braće Branchetta 20, 51000 Rijeka, Croatia

ABSTRACT

The use of synthetic data in biomedicine is growing rapidly, offering an alternative when real patient data is scarce or restricted due to privacy concerns. Synthetic datasets aim to replicate the statistical properties of real data, enabling researchers to develop models and perform analyses without compromising confidentiality. However, their application raises methodological concerns—particularly regarding when and how synthetic data should be used in the analysis pipeline. A key issue is whether analyses should be conducted on real data and validated with synthetic data, or whether synthetic data can serve as the primary basis for analysis. The lack of consensus poses questions about the reliability of findings derived solely from synthetic datasets. This study explores the issue using Tabular Variational Autoencoder (TVAE) to generate synthetic versions of a bladder cancer recurrence dataset. The authors compare correlation and feature importance results from synthetic and original data. Findings show that while synthetic data can reflect broad trends, model sensitivity—especially with Random Forests—can lead to discrepancies in feature importance and predictive accuracy. In contrast, basic statistical methods are more stable. These results highlight the need for careful methodological planning and transparent reporting when using synthetic data, as analytical outcomes may depend heavily on when and how such data is introduced.

Keywords: Biomedical data analysis, Feature importance evaluation, Machine learning, Synthetic data, Tabular variational autoencoder

INTRODUCTION

Data synthetization is an extremely useful tool in data science, especially biomedicine where limited amounts of data are common due to well-known difficulties in data acquisition within this domain (Zheng, 2015). While data synthetization has many potential benefits – such as data pseudo-anonymization, creation of more robust machine learning models and overall model quality increase, an important note remains that this data is not real-world data, and the methods used for synthetization may or may not introduce changes to the data that may not be readily apparent with common analysis techniques (Giles, 2022).

One of the techniques that is commonly used for data analysis is feature importance determination. While this technique is commonly used by machine learning experts as a first step in feature engineering approach, it can also be used as a technique to determine which of the parameters influence the output variable – e.g. which of the patient metrics influence the likelihood of a positive diagnosis (Baressi Šegota, 2024).

A question arises – should the analysis of this be performed only on original data, or can it be done on synthetic data? Most synthetization techniques are designed to copy the statistical distributions within data, which may not copy the exact feature influences as well. This paper tests one of the most commonly used techniques – tabular variable autoencoder (TVAE) on a bladder cancer recurrence dataset, and evaluates the connection between variables with four feature importance metrics in order to determine if there is a change of feature importances between original and synthetized datasets. The paper will first present the used methodology – including the used dataset, synthetization approach and feature importance determination. Following this, the results of feature scores will be presented and discuss, with the final part presenting the conclusions drawn from those results.

METHODOLOGY

Dataset

The dataset used in this study is the Bladder Cancer Recurrence dataset (Singh, 2021) - a freely publicly available dataset consisting of data for 294 patients. The data used as the input in this study consists of three data vectors – the number of individual tumors, their size (in centimeters, for largest tumor) and the type of treatmant applied (placebo, pyridoxine or thiotepa – encoded as numerical classes 0, 1 or 2 respectively). The targeted output was a binary target indicating whether a recurrence of the cancer occurred within the patient (0 indicating no recurrence, and 1 indicating one or more recurrences) (Andrews, 1985; Wei, 1989).

Data Synthetization With TVAE

The Tabular Variational Autoencoder (TVAE) is a generative model designed to synthesize tabular data while preserving statistical relationships among variables (Li, 2019). It extends the conventional Variational Autoencoder (VAE) framework to accommodate heterogeneous data types—continuous, ordinal, and categorical—commonly found in structured datasets (Öğretir, 2022). The TVAE comprises two neural networks: an encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and a decoder $p_\theta(\mathbf{x}|\mathbf{z})$, where $\mathbf{x} \in \mathbb{R}^d$ denotes a data sample in d-dimensional space, and $\mathbf{z} \in \mathbb{R}^k$ is a latent variable sampled from a prior distribution. The encoder maps the observed data $\mathbf{x}$ to a latent representation $\mathbf{z}$ by approximating the posterior distribution, while the decoder reconstructs $\mathbf{x}$ from $\mathbf{z}$. A standard multivariate Gaussian prior is assumed on the latent space (Tan, 2024):

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{1}$$

The objective function of the TVAE is the evidence lower bound (ELBO), which balances reconstruction accuracy and regularization of the latent space (Fonseca, 2023):

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\right) \tag{2}$$

Here, D_{KL} denotes the Kullback-Leibler divergence between the approximate posterior and the prior distribution. The reconstruction term is modeled using appropriate likelihood functions based on data types: Gaussian for continuous variables and categorical cross-entropy for discrete variables. To stabilize training, TVAE discretizes categorical features using one-hot encoding and normalizes continuous features via Gaussian Mixture Models (GMMs), preserving multimodal distributions (Apellániz, 2024). Once trained, synthetic data generation proceeds by sampling latent vectors $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and passing them through the decoder (Wu, 2024):

$$\widehat{\mathbf{x}}_i = \underset{\mathbf{x}}{\mathrm{argmax}}\, p_\theta(\mathbf{x}|\mathbf{z}_i) \tag{3}$$

The decoder network outputs parameters for each feature's distribution, from which synthetic values are sampled. For categorical variables, this corresponds to drawing from a softmax distribution; for continuous ones, from a parameterized Gaussian mixture. TVAE's capacity to learn complex joint distributions makes it particularly suited for applications requiring realistic data synthesis in structured domains, such as biomedical datasets. Importantly, by training the model on real patient records and subsequently generating new records from latent samples, researchers can perform downstream tasks (e.g., correlation analysis, classification, or feature ranking) without direct exposure to sensitive data, thereby enhancing privacy without entirely sacrificing fidelity (Apellániz, 2024). Nonetheless, like all generative models, the quality of the synthetic data depends on latent dimensionality, network architecture, and data preprocessing pipelines. Empirical validation remains essential to confirm the preservation of meaningful relationships and avoid artifacts that may distort downstream inferences (Tan, 2024).

In this study, the model is developed with training for 25,000 epochs, which achieved the overall data similarity index of 94,56%, indicative of a high-quality data synthetization. A total of 1000 points are synthetized and mixed with the data at different levels (100, 250, 500, 750 and 1000 points), prior the feature influence analysis.

Feature Influence

Identifying relevant features is a central task in biomedical data analysis, particularly when working with high-dimensional datasets or synthetic data. Two complementary strategies for quantifying feature importance are examined in this study: statistical correlation-based approaches, and model-based importance estimation using ensemble learning. Correlation methods quantify the degree of association between individual features and a target variable, typically assumed to be continuous or ordinal. Let $x_i \in \mathbb{R}^n$ denote the vector of observations for feature i, and $y \in \mathbb{R}^n$ the corresponding target values. The Pearson correlation coefficient evaluates the strength of a linear relationship between x_i and y, assuming both variables are normally distributed (Benesty, 2009):

$$\rho_{\text{Pearson}}\left(x_i, y\right) = \frac{\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)(y_j - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2}\sqrt{\sum_{j=1}^{n}(y_j - \bar{y})^2}} \tag{4}$$

This metric captures only linear dependencies and is sensitive to outliers and non-normality.

The Spearman rank correlation assesses monotonic relationships by comparing ranks rather than raw values. If $R(x_i)$ and $R(y)$ are the rank-transformed vectors (Wissler, 1905):

$$\rho_{\text{Spearman}}\left(x_i, y\right) = \rho_{\text{Pearson}}\left(R\left(x_i\right), R(y)\right) \tag{5}$$

Spearman correlation is robust to outliers and capable of detecting non-linear monotonic trends.

Kendall's τ measures ordinal association by evaluating the concordance of pairwise observations (Schaeffer, 1956):

$$\tau\left(x_i, y\right) = \frac{C - D}{\binom{n}{2}} \tag{6}$$

where C is the number of concordant pairs and D is the number of discordant pairs. Kendall's τ is particularly suitable when the data exhibit ties or a high degree of ordinal noise. In all three methods, the magnitude of the correlation coefficient is taken as a proxy for feature importance. Features with coefficients near zero are assumed to contribute little explanatory power to the target variable.

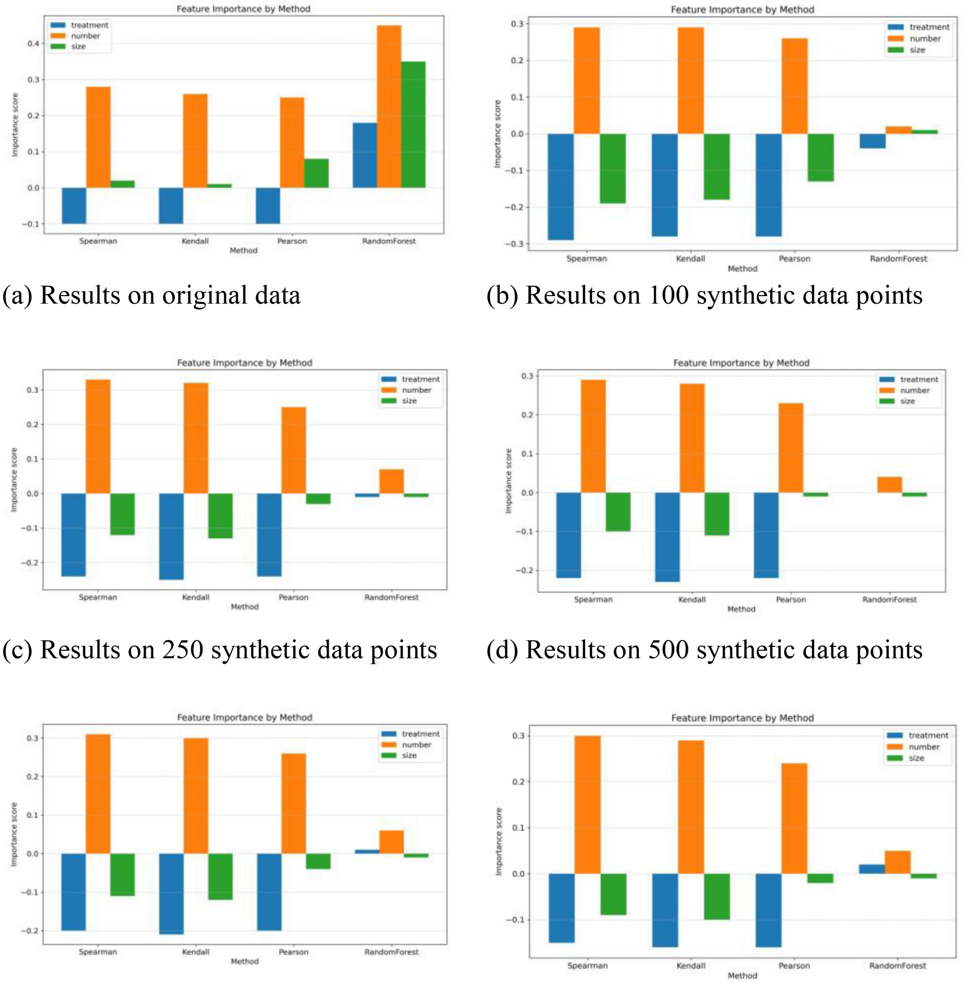
Random Forests (RF) are ensemble learning models based on collections of decision trees trained on bootstrap samples and random feature subsets (Tanha, 2017). Feature importance in RF models is typically computed using the mean decrease in impurity (MDI), which quantifies how much each feature reduces the Gini impurity or mean squared error across all trees in the forest. Let T denote the set of all decision trees in the ensemble, and for each feature x_i, let $\mathcal{S}_i^t$ be the set of splits involving x_i in tree $t \in T$. The total importance of x_i is (Šegota, 2025):

$$\text{Imp}\left(x_i\right) = \frac{1}{|T|}\sum_{t \in T}\sum_{s \in \mathcal{S}_i^t} \Delta\mathcal{I}(s) \cdot \frac{N_s}{N_t} \tag{7}$$

where $\Delta\mathcal{I}(s)$ is the impurity reduction at split s, N_s is the number of samples reaching that node, and N_t is the total number of samples in tree t. This method captures non-linear interactions and is robust to multicollinearity among features. However, it may exhibit bias in favor of features with more levels or broader numerical ranges (Meyer, 2021). In the current study, feature rankings obtained from correlation-based and model-based analyses are compared to assess the fidelity of synthetic data in preserving variable relevance. Discrepancies in these rankings serve as indicators of structural distortion introduced during the synthetization process.

RESULTS AND DISCUSSION

The data shown in Figure 1. demonstrates the scores of various metrics, per each of the targets. The first given graph is the feature importance on the original data, while the remaining ones are evaluations on the synthetic datasets with different amount of data points. While some of the influences are contained, we can see changes in the correlation smaller in case of number of tumors, which remains similar, but higher for other two metrics. Notably, the size metric reverses its correlation into negative. This change is also the case for the random forest feature importance, with the metrics showing a significantly different scores compared to the original dataset.



(a) Results on original data

(b) Results on 100 synthetic data points

(c) Results on 250 synthetic data points

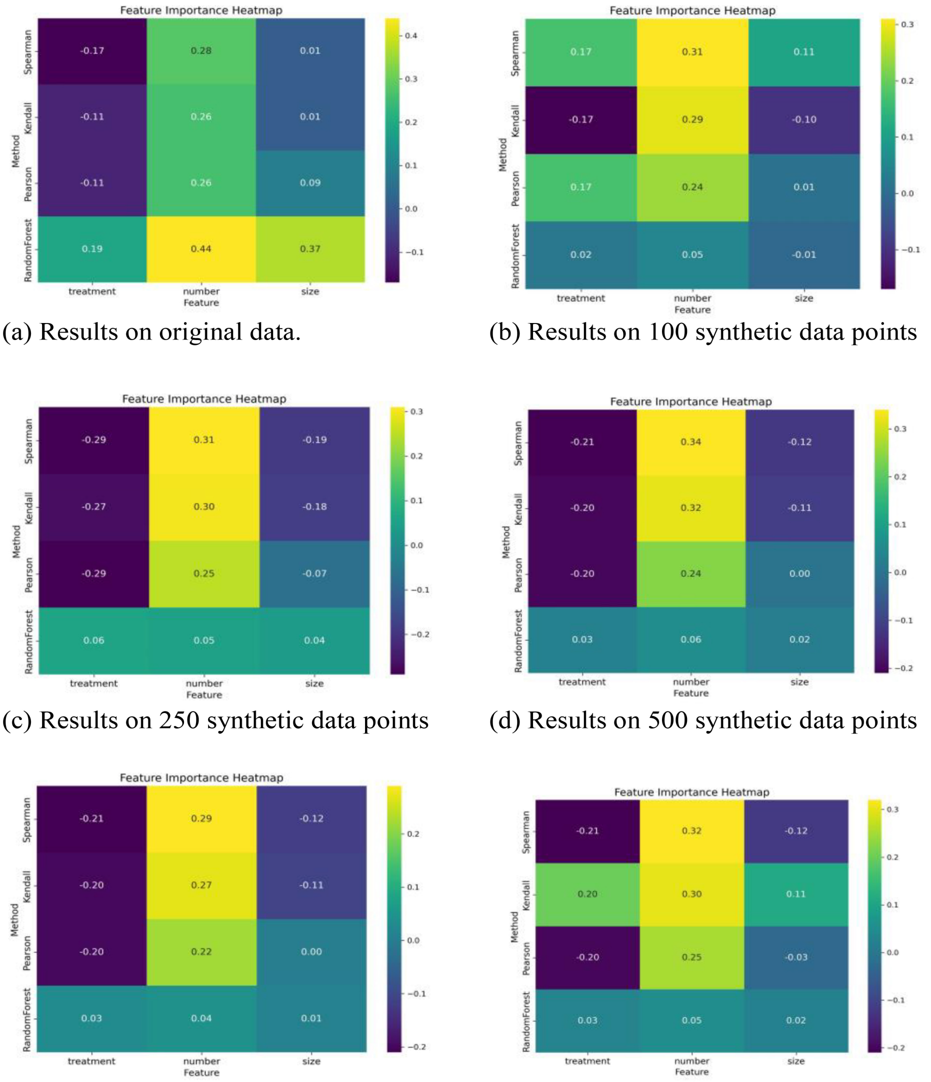
(d) Results on 500 synthetic data points

(e) Results on 750 synthetic data points

(f) Results on 1000 synthetic data points

Figure 1: The comparison of metrics on real and synthetic datasets.

The similar can be seen in Figure 2, where the scores are shown as heatmaps. Each row presents one of the evaluation methods, and each column presents one of the used inputs. The position of the images is the same, with the first (upper left) being the original data heatmap, while the others are calculated on the different amounts of synthetic data. This visualization confirms the previously given one with a clear difference visible between the heatmap for original data and synthetic data. The differences between the different amounts of sznthetic data are present, but not not as pronounced. The differences seen between these subfigures in Figure 2 are expected statistical variation introduced by undersampling data.



(a) Results on original data.

(b) Results on 100 synthetic data points

(c) Results on 250 synthetic data points

(d) Results on 500 synthetic data points

(e) Results on 750 synthetic data points

(f) Results on 1000 synthetic data points

Figure 2: The comparison of heatmaps of scores.

CONCLUSION

This paper performed the analysis with the goal of testing how the feature influence changes between the original, real-world collected, data and the data that was generated using TVAE synthesizer. The dataset used dealt with the bladder cancer recurrence rates – with three features tested (size of tumor, number of tumors and the type of used therapy. A total of 1000 data points are synthesized, and the evaluation is performed on different sized, randomly selected, subsets of the data.

The results demonstrate significant differences in the measured feature importances between original data and the synthesized data. While the values are consistent amongst the different sized synthetic datasets, they differ greatly from the original data. This points towards an important conclusion and that is that even if the synthetization method shows high performance (standard evaluation techniques – comparison of column shapes and column data pairs shown a high evaluation of over 94% for the used synthetic dataset), it cannot be assumed that the feature influences contained in the synthesized dataset are going to be the same as in the original data, at least within the confines of the presented experimental setup. This means that any feature influence analysis should be performed on original data, and not synthesized.

Future work should focus on the generalization of the presented research – testing different synthetization techniques on different datasets to see if there are synthetization methods that allow for the feature importance to be tested on the synthesized datasets.

ACKNOWLEDGMENT

This research is supported by SPIN project “INFOBIP Konverzacijski Order Management” (IP.1.1.03.0120), “Projektiranje i razvoj nove generacije laboratorijskog informacijskog sustava (iLIS)” (IP.1.1.03.0158), “Istraživanje i razvoj inovativnog sustava preporuka za napredno gostoprimstvo u turizmu (InnovateStay)” (IP.1.1.03.0039), “European Digital Innovation Hub Adriatic Croatia (EDIH Adria) (project no. 101083838)” under the European Commission’s Digital Europe Programme and the FIPU project “Sustav za modeliranje i provedbu poslovnih procesa u heterogenom i decentraliziranom računalnom sustavu”, and University of Rijeka research grants UNIRI-IZ-25-220 and UNIRI-IZ-25-6.

REFERENCES

Andrews, D. F., & Hertzberg, A. M. (1985). DATA: A collection of problems from many fields for the student and research worker. Springer-Verlag.

Apellániz, P. A., Parras, J., & Zazo, S. (2024, August). An improved tabular data generator with VAE-GMM integration. In 2024 32nd European Signal Processing Conference (EUSIPCO) (pp. 1886–1890). IEEE.

Baressi Šegota, S. (2025). Determining the energy-optimal path of six-axis industrial robotic manipulators using machine learning and memetic algorithms (Doctoral dissertation, University of Rijeka. Faculty of Engineering).

Baressi Šegota, S., Anđelić, N., Štifanić, J., Štifanić, D., & Car, Z. (2024). Influence of dimensionality reduction approaches on various machine learning models for a biomedical high-dimension dataset. In Book of Proceedings 3rd Serbian International Conference on Applied Artificial Intelligence (SICAAI). Kragujevac: University of Kragujevac.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In Noise reduction in speech processing (pp. 1–4). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fonseca, J., & Bacao, F. (2023). Tabular and latent space synthetic data generation: A literature review. Journal of Big Data, 10(1), 115.

Giles, O., Hosseini, K., Mingas, G., Strickson, O., Bowler, L., Smith, C. R.,... & Vollmerteke, S. (2022). Faking feature importance: A cautionary tale on the use of differentially-private synthetic data. arXiv preprint arXiv:2203.01363.

Li, S. C., Tai, B. C., & Huang, Y. (2019, December). Evaluating variational autoencoder as a private data release mechanism for tabular data. In 2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC) (pp. 198–1988). IEEE.

Meyer, A., Albarghouthi, A., & D'Antoni, L. (2021). Certifying robustness to programmable data bias in decision trees. Advances in Neural Information Processing Systems, 34, 26276–26288.

Öğretir, M., Ramchandran, S., Papatheodorou, D., & Lähdesmäki, H. (2022, December). A variational autoencoder for heterogeneous temporal and longitudinal data. In 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1522–1529). IEEE.

Schaeffer, M. S., & Levitt, E. E. (1956). Concerning Kendall's tau, a nonparametric correlation coefficient. Psychological Bulletin, 53(4), 338.

Singh, U. [utkarshx27]. (2021). Bladder cancer recurrence dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/utkarshx27/bladder-cancer-recurrences

Tan, Y., Zhu, H., Wu, J., & Chai, H. (2024). DPTVAE: Data-driven prior-based tabular variational autoencoder for credit data synthesizing. Expert Systems with Applications, 241, 122071.

Tanha, J., Van Someren, M., & Afsarmanesh, H. (2017). Semi-supervised self-training for decision tree classifiers. International Journal of Machine Learning and Cybernetics, 8(1), 355–370.

Wei, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American Statistical Association, 84(408), 1065–1073.

Wissler, C. (1905). The Spearman correlation formula. Science, 22(558), 309–311.

Wu, J., Chen, S., Zhao, Q., Sergazinov, R., Li, C., Liu, S.,... & Brunzell, H. (2024, March). Switchtab: Switched autoencoders are effective tabular learners. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(14), 15924–15933.

Zheng, Y. (2015). Methodologies for cross-domain data fusion: An overview. IEEE Transactions on Big Data, 1(1), 16–34.