Constraint-Aware Tabular Variational AutoEncoder for Synthetic Data Generation in Health Domain

1 Abstract

Background and motivation Healthcare AI growth in resource-constrained regions, particularly in Africa faces majors challenges including limited access to quality dataset, privacy regulation and inadequate data sharing infrastructures. This constitute a critical gap for data-driven healthcare innovation. Synthetic data generation growths as a promising solution, yet existing approach like Tabular Variational AutoEncoder(TVAE) focus primary on statistical similarity between real data and synthetic data without including domain-specific logical consistency resulting on statistical similar data but medically irrelevant.

Methodology We propose Constraint-Aware TVAE (CA-TVAE), an extension of traditional TVAE that incorporates domain-specific constraints through post-processing validation and correction. Our approach defines constraints as logical functions $C = \{ c_1, c_2, ... c_n\}$ where $c_i(\mathrm{x}) = 1$ indicates constraint satisfaction. We implemented eight cardiovascular domain-specific relational constraints including: Blood Pressure Grade Constraints, Diabetes-Related Constraints, Lifestyle Consistency, Socioeconomic Logic. The methodology involves training a standard TVAE model, generating synthetic data, validating against defined constraints, and applying corrective post-processing to ensure constraint satisfaction.

Experiments We evaluated our approach using cardiovascular survey data from northern Senegal (n=911, 168 features after preprocessing). Statistical fidelity was assessed using Kolmogorov-Smirnov tests for numerical features and χ^2 tests for categorical features. Machine learning utility was evaluated using the Train-Synthetic-Test-Real (TSTR) framework with Random Forest classifiers.

Key Results Statistical Fidelity: 26/43 numerical features and 100/125 categorical features passed distribution similarity tests ML Utility: CA-TVAE achieved 97.1% AUC compared to 97.4% on original data and 94.2% with standard TVAE. Constraint Satisfaction: 100% compliance with all defined domain constraints post-processing

Impact and Significance This work addresses critical healthcare data challenges in under-represented regions by providing a practical framework for

generating medically consistent synthetic data. The constraint-aware approach ensures that synthetic cardiovascular data maintains not only statistical properties but also clinical logic, making it suitable for training robust AI models. This is particularly valuable for African healthcare systems where data scarcity significantly limit AI development. Our approach demonstrates that domain expertise can be effectively integrated into synthetic data generation, paving the way for more reliable AI applications in resource-constrained healthcare environments. The methodology is generalizable to other medical domains and geographical contexts where similar data challenges exist.

Future Directions Future work will focus on automating constraint discovery, expanding evaluation across multiple healthcare datasets, and developing more sophisticated constraint integration mechanisms within the TVAE architecture rather than through post-processing alone.

Keywords Artificial Intelligence, Cardio-vascular disease, Machine Learning, Predictors

2 Introduction

Since a while, AI researchers mostly focused on model performance by design the best architecture, tuning hyper-parameters or choosing the best model but with the enormous researches made by both AI researchers and other stakeholders in AI-specific context, it therefore follows that the enhancement of AI systems can be yield by focused on dataset quality instead of model quality. We moved from Model-Centric AI to Data-Centric AI, as right now focusing on data quality can be more benefic and will fulfil model's performance [1]. This demand for a comprehensive data is crucial for the development of reliable and robust AI-model for healthcare

However health data may be incomplete, difficult to collect, not generalised or governed by privacy standards that are problematic to uphold. Collecting high-quality medical data is time consuming and resource-intensive, whereas it's essential to ensure effective AI models [2]. In Africa, developing effective AI systems for healthcare is seriously constraining by a multitude of factors. Limited access of datasets, not digitalized health facilities, poor data sharing regime, privacy laws[3]. Quality data is a essential to disease understanding, pathology detection, early diagnosis to reduce complications and mortality,personalized treatment strategies, and a plethora of applications. The tension between the need for data-driven medical innovation and the right to patient privacy is a significant bottleneck in healthcare research. This is not just a technical challenge, but a complex socio-legal one, which highlights the urgent need for innovative solutions to bridge this critical gap. Using synthetic data can be really hard in this condition but it can consist of an innovative and effective approach to reach actionable models to avoid being left behind concerning technological advance on AI in healthcare.

In response to this limitations, researchers developed a variety of synthetic data generation techniques and considered it as a reliable and promising alternative. Synthetic data generation claims to create artificial data that mimic the original real data by reproducing the statistic and distribution properties of the real-world dataset. such as Bayesian models [4], Variational AutoEncoders(VAE) [5] or Generative Adversarial Networks(GAN) [6] that can ensure effective generation of medical data. GAN models

and VAE have proved to be very efficient when it comes to generating medical data, such as X-rays and patient medical records. In the other hand Tabular VAE, a synthetic data approach based on VAE specifically designed for tabular data structures commonly found in healthcare datasets demonstrate very good performance on medical data such as data from Electronic Medical Records(EMR). Traditional TVAE focused on statistical similarity without constraint awareness leading to synthetic data that are statistically plausible but might be logically incorrect depending on the domain. This limitation is lead us to

3 Methodology

Problem formulation

Given a cardiovascular dataset D = { x_1,x_2,...x_n } where each sample x_i $\mathbb{R}^d$ contains both numerical and categorical features, we aimed to generate data D_{synth} that:

- Maintains the statistical fidelity of the original distribution of real features
- Satisfy the domain-specific constraints C = { c_1, c_2,... c_n }
- Enables predictive ML tasks

Constraint-Aware Extension

Our CA-TVAE extends the baseline architecture with constraint-aware consideration: Constraint Specification: We define constraints C as a set of functions

c_i: X $\rightarrow$ 0,1 where $c_i(x) = 1$ indicates constraint satisfaction. Constraint is define with:

Relational constraints: c_rel(x) = 1 if logical relationship holds. In the case of the explored dataset, relational constraints are important to keep the logic between features. We implemented constraints like:

- BPGradeConstraint define as: When $'Diastolique >= 90'$ and $'Systolique >= 140'$ equal 0 then the BPGrade is equal to '0'
- trconduction
- InsulineandADOConstraint: when the variable $'Diabetic'$ equals 0, $'Insuline'$ and $'ADO'$ should equal 0
- TypeLeftVentricularHypertrophyconstraint When the Left Ventricular Hypertrophy(LVH) is equals to 0 then the TypeLVH must be "NON EXISTENT".
- SourceOfIncomeConstraint: when the $'Monthlyincome'$ equals to 0 then the $'Incomesource'$ should be $'NONEXISTENT'$
- ReasonNoSportConstraint: when the $'Sedentarity'$ is not 0 then $'ReasonNoSport'$ should be $'NONEXISTENT'$
- StockandSaltAssociationConstraint: when the $'StockConsumption'$ is 0 then $'Quantityofstock'$ and the $'StockandSaltAssociation'$ should be 0.
- QuantityFruitAndFruitConsumedConstraint: when $'FrequencyFruitConsumption'$ equals to 0 then the $'QuantityFruit'$ should be 0 and the $'ConsumedFruit'$ should be $'NONEXISTENT'$

Data generation techniques for tabular data

Tabular Generative Adversarial Networks (GANs) [7]and Variational Autoencoders (VAEs) [8] have been extensively applied to tabular data synthesis. CTGAN and TableGAN adapt adversarial training for tabular data, while TVAE [8] leverages the stable training properties of VAEs with specialized encoders for mixed-type tabular data.

3.1 Tabular Variational AutoEncoder

TVAE employs a variational autoencoder based neurone network to train data, it was specially design for mixed-type tabular data. The architecture of Tabular VAE can be breaking down into two key models.

- The encoder takes the original data and compress it into a lower-dimensional representation known as latent space, it learns to map the input to a probability distribution within the latent space
- The decoder's function is to take data from the latent space and tries to restore the relationship between the different variables.

TVAE loss function is build with two principal components: the reconstruction loss and the KL Divergence loss. The reconstruction loss aims to evaluate the decoder capacity to reconstruct the data from the latent representation ensuring that the generated data is well represented the input. The KL Divergence loss is supposed to measure the difference between the learned latent distribution and the prior distribution. Minimizing the KL Divergence loss, encourage the model to create smooth and continuous realistic latent space essential for creating realistic data. We proposed an synthesizer based on TVAE with added constraint after generation, the process of our methodology is to train a TVAE model with processed data, verify if the generated data satisfy defined constraints. If there is any unsatisfied constraint, the corresponding constraint is then added as describe in 1. Generated data was evaluate for statistical fidelity, machine learning utility compare to real data and data generated with standard TVAE.

=0

4 Experiments

Dataset

Data were gathered from a survey conducted in northern Senegal between November and December 2012. The collected data contains 1017 observations and 224 socio-demographic and clinical variables to examine cardiovascular dysfunctions. The database contains many missing data. The target feature, Exam Cardiovascular (EC for short) is the result of the cardiovascular examination: it represents the final global cardiovascular result. The target feature is bivariate, with class 0, which means an absence of cardiovascular dysfunction and class 1 corresponds to a presence of cardiovascular dysfunction.

Data Cleaning

Removing unused columns Ultra Null columns: 20 Certain variables in the data set contain more than 1000 missing values, and these variables can cause

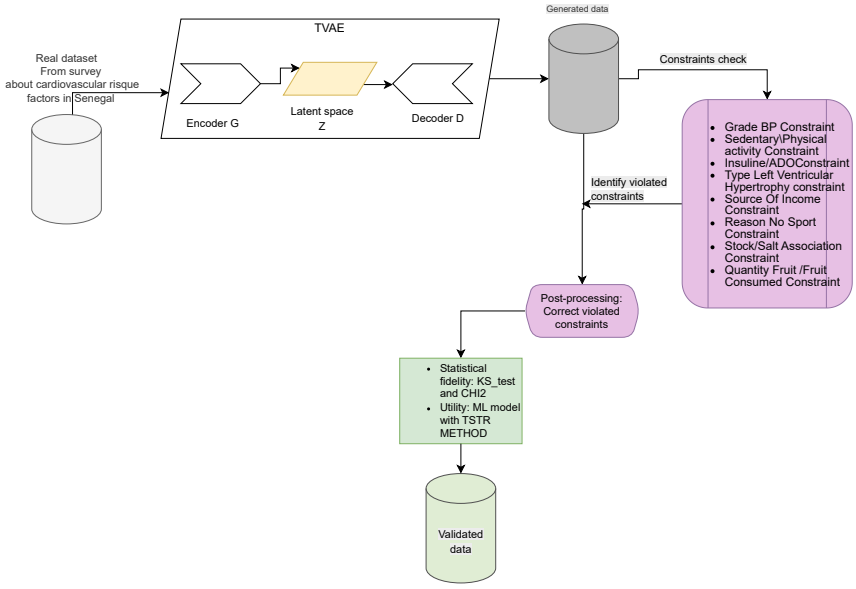

Figure 1: Proposed methodology for constraint added

mis-analysis when handling their missing values. In our case, we decided not to include them in the analysis process, so about 20 columns were removed. Some other variables are related to treatment and contains also many missing values, as we are not conducting an analysis about the treatment of cardiovascular disease, they were removed.

Data transformation

Categorical features mostly contain more than hundred categories and this need to be handled in order to limited them. We limited the number of category in some variables like $AlimentMatin$: 373 categories, $Fruitconsom$: 183 categories $AlimMidi$: 138 categories $Alimsoir$: 476 categories to 9 categories as the variables related to 'frequency' have 9 categories. The first 8 categories are maintained and the other will be group in one category named 'Other'.

Handle missing values thought variable dependency: The first four letters of the column name are the code for that column. When there is a dependence between variables, the code is usually the same for the whole group. We create a group of features depending on the code in order to handle missing values through dependence. For example, "N016" is the code for features related to alcohol. When the "alcohol consumption" column is 0, all the other features in the group are 0 or "NON EXISTENT", meaning that they did not exist. The variable $BPGrade$ depends on $Diastolique >= 90$ and $'Systolique >= 140'$, and when these two columns equal 0, the dependent column also equals "0". $BPGrade$ corresponds to the level of high

blood pressure.

Handling remained missing values Remained missing values was handled with the column mode for categorical characteristics and the mean for the numerical variables.

Results

Statistical comparison of real and synthetic data: Fidelity is about how closely generated samples mirror the distributional properties and structural characteristics of the original dataset. When dealing with mixed-type dataset composed with both continuous numerical features and categorical features, it's crucial to establish a combination of statistic tests to assess the fidelity comprehension of original or excepted data. In our pipeline, we combine two relevant statistical tests, Kolmogorov-Smirnov and $\tilde{\chi}^2$ test as they are considered as the common used methods [9] to ensure that both numerical values and categorical labels maintain their integrity as the new data is generated, assuming a high level of confidence in the overall data quality. We choose a p-value equals to 0.05 to be able to reject the null hypothesise in the two performed tests. For numerical continuous features, the Kolmogorov-Smirnov (KS-test) test is a optimal approach for verifying the reliability of the characteristics. The KS-test is a non-parametric test used to determine if two samples of data come from the same distribution. It works by calculating the empirical distribution of all the numerical features in the real dataset and compare them to the corresponding ones calculating in the generated data.

$$D_{n,m} = \sup_{x \in \mathbb{R}} |F_n(x) - G_m(x)| \tag{1}$$

Where:
$F_n(x)$ is the empirical distribution function of the first sample (size n)
$G_m(x)$ is the empirical distribution function of the second sample (size m)

For the categorical features, the chi-square $\tilde{\chi}^2$ test was used, $\tilde{\chi}^2$ is a statistical method that is used when the features to be considered are categorical ones. Its purpose is to evaluate the relationship between the variables statistically, and it is often used in clinical research. In the case of fidelity test, we will perform a $\tilde{\chi}^2$ goodness of fit test to check the match between categorical features in real data and generated ones. We resumed the statical quality of the generated data in table 1

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

Where:
O_i is the observed frequencies in category i
E_i is the expected or original frequencies in category i
k is the number of categories

Machine learning model trained on synthetic data can be generalize on real situation: Unlike fidelity, the utility test of the generated data is not just a statistical measurement of the data but it aims to ensure that the created dataset can serve as a viable and safe substitute for real, sensitive data by its predictive power.

	Numerical features(ks-test)	categorical features(χ^2)
Real data(168)	43	125
Synthetic matched features	26	100

Table 1: Summary comparison between synthetic and real data

Model	Random Forest(AUC)
Original data	97.4%
Standard TVAE	94.2%
TVAE-CA(post-processing added cons)	97.1%

Table 2: Synthetic data performance over real data performance on machine learning model

The utility of the synthetic data can be measure with a machine learning model trained on it, if the model perform comparably on the synthetic and real data, the high utility of the synthetic data can be assume. We sought to demonstrate that generated data are as relevant that the real data for a machine learning model, we then used the Train-Synthesize Test-Real(TSTR)[10] method which is a widely used approach to ensure the generalization of synthetic data to real situations.

We trained Random Forest(RF) model with 80% of the real data and tested it on 20% left, we then trained the same model with 80% of the synthetic data and tested it on the real test set. The results of our approach was evaluated thought the Area Under Curve(AUC) metric which is resumed in table2. Training RF model on original dataset, generated data with TVAE with constraints added during training and generated data with TVAE-CA. We noticed that RF performed as well as with TVAE-CA than original dataset unlike the standard TVAE regrading the AUC metric attending respectively 97.4%,97.1.%, and 94.2%.

References

[1] E. Y. Chang, "Knowledge-Guided Data-Centric AI in Healthcare: Progress, Shortcomings, and Future Directions," Apr. 2023.

[2] C. S. Kwok, E.-A. Muntean, C. D. Mallen, and J. A. Borovac, "Data Collection Theory in Healthcare Research: The Minimum Dataset in Quantitative Studies," *Clinics and Practice*, vol. 12, pp. 832–844, Oct. 2022.

[3] K. Gulma, Z. Saidu, K. Godfrey, A. Wada, Z. Shitu, A. Bala, S. B. Borodo, S. Julde, and M. Mohammed, "Harnessing Machine Learning for Predictive Healthcare: A Path to Efficient Health Systems in Africa," *Health Informatics and Information Management*, May 2025.

[4] G. Gogoshin, S. Branciamore, and A. S. Rodin, "Synthetic data generation with probabilistic Bayesian Networks," *Mathematical biosciences and engineering: MBE*, vol. 18, pp. 8603–8621, Oct. 2021.

[5] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," June 2014.

[7] L. Xu and K. Veeramachaneni, "Synthesizing Tabular Data using Generative Adversarial Networks," Nov. 2018.

[8] H. Ishfaq, A. Hoogi, and D. Rubin, "TVAE: Triplet-Based Variational Autoencoder using Metric Learning," Feb. 2023.

[9] V. Hudovernik, M. Jurkovič, and E. Štrumbelj, "Benchmarking the Fidelity and Utility of Synthetic Relational Data," Oct. 2024.

[10] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, and C. S. Greene, "Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing," *Circulation. Cardiovascular Quality and Outcomes*, vol. 12, p. e005122, July 2019.