# Supplementary Material: Frugal Generative Modeling for Tabular Data

Alice Lacan[1,2] (✉), Blaise Hanczar[1], and Michele Sebag[2]

[1] IBISC, U. Evry, Université Paris-Saclay
{alice.lacan,blaise.hanczar}@univ-evry.fr
[2] TAU, CNRS-INRIA-LISN, Université Paris-Saclay Michele.Sebag@lri.fr

These supplementary materials complement the paper "Frugal Generative Modeling for Tabular Data" with additional quantitative and qualitative results, as well as reproducibility information and settings. They are organized as follows:

## A. Benchmark datasets

*2d toy datasets.* The toy datasets used in our experiments can be easily retrieved from the scikit-learn library: scikit-learn.org/stable/api/sklearn.datasets.html.

*Medium-size datasets.* We detail the references for the medium-size open-source ML tabular datasets hereafter:

- Pima Indians Diabetes Database: https://www.openml.org/d/37
- Gesture Phase Prediction: https://archive.ics.uci.edu/dataset/302/
- Magic Gamma Telescope: https://archive.ics.uci.edu/dataset/159/
- Wilt dataset: https://www.openml.org/d/40983

The preprocessing of these datasets involves only the standardization to zero mean and unit variance and the one-hot encoding of the target variables for the conditioning of GMDA .

*Higher dimensions.* We also trained GMDA on the high-dimensional GTEx and TCGA gene expression datasets. TCGA data was retrieved using the RTCGA[3] package in R (release date 2016-01-28), while the latest version (v8) of GTEx data (with TPM normalization) was obtained from the open-access portal [4]. Since both datasets include circa 20,000 coding genes, we reduced the number of dimensions to circa 1,000 genes using only landmark genes in the first instance. The preprocessing procedure, involving removal of duplicates, landmark genes IDs mapping, quantile transformation and standardization, can be found in our open-source code. For GTEx, we end up with 974 variables and 26 target tissues to classify (respectively 978 variables and 24 tissues for TCGA).

[3] https://bioconductor.org/packages/RTCGA.

[4] https://gtexportal.org/home/downloads/adult-gtex

## B. Best settings

### 1. Baseline TVAE, CTGAN and TabDDPM

For the sake of reproducibility, we detail the hyper-parameters used for each baseline model and each dataset in this section. Such hyper-parameters were selected based upon the original papers of TVAE, CTGAN and TabDDPM.

Table 1: Best hyper-parameters of TVAE.

| Hyper-parameter | Diabetes | Gesture | Magic | Wilt |
|---|---|---|---|---|
| Layers | [512, 512, 512, 64] | [512, 256] | [256,512,512,128] | [256,512,512,128] |
| Batch size | 256 | 256 | 256 | 2048 |
| Learning rate | 5.7e-2 | 4.7e-4 | 1.3e-3 | 1.3e-3 |
| Optimizer | Adam | Adam | Adam | Adam |
| Loss factor | 6.32 | 2.43 | 6.54 | 6.54 |
| Weight decay | .0 | .0 | .0 | 1e-6 |
| Epochs | 20,000 | 30,000 | 5,000 | 5,000 |

Table 2: Best hyper-parameters of CTGAN.

| Hyper-parameter | Diabetes | Gesture | Magic | Wilt |
|---|---|---|---|---|
| Layers | [128, 512,512, 512,512,64] | [256,256, 256, 128] | [512,128, 128,512] | [512,128, 128,512] |
| Batch size | 256 | 2048 | 2048 | 2048 |
| Learning rate disc. | 2e-4 | 2e-4 | 1.3e-3 | 1.3e-3 |
| Learning rate gen. | 2e-4 | 2e-4 | 2e-4 | 2e-4 |
| Optimizer | Adam | Adam | Adam | Adam |
| Loss factor | 6.32 | 2.43 | 6.54 | 6.54 |
| Epochs | 10,000 | 30,000 | 5,000 | 5,000 |

### 2. Baseline VAE and WGAN-GP

In the interest of time and resources, we limited the search for the best hyper-parameters to the batch size, learning rate, optimizer, and hidden layers dimensions for the VAE and WGAN-GP over 1000 epochs. As numerical errors can arise with the VAE, it is useful to force the initialization of weights to minimal values ( 8.10-2). As regards the WGAN-GP, the discriminator requires five additional iterations than the generator, and each network needs its own learning

Table 3: Best hyper-parameters of TabDDPM.

| Hyper-parameter | Diabetes | Gesture | Magic | Wilt |
|---|---|---|---|---|
| Input transfo. | Quantile | Quantile | Quantile | Quantile |
| Layers | [128,512] | [128,512,512,1024] | [1024,2048, 2048,1024] | [1024,512,512,512, 512,512,512,128] |
| Batch size | 491 | 4096 | 4096 | 256 |
| Learning rate | 1.15e-5 | 2.8e-3 | 1.8e-3 | 1.3e-4 |
| Optimizer | AdamW | AdamW | AdamW | AdamW |
| Loss factor | 6.32 | 2.43 | 6.54 | 6.54 |
| Timesteps | 1,000 | 1,000 | 1,000 | 100 |
| Epochs | 30,000 | 30,000 | 100,000 | 30,000 |

rate so that none of them outperforms the other too quickly. The training time is 1 hour and 40 minutes for the best WGAN-GP on TCGA (batch size is 64), and 21 minutes on GTEx (batch size is 1024).

Table 4: Best hyperparameters for the VAEs trained on landmark genes.

| Hyper-parameter | GTEx | TCGA |
|---:|---:|---:|
| Batch size | 256 | 2048 |
| Layers | [8192, 4096, 4096, 2048] | [2048, 8192, 256, 4096] |
| Latent dim | 128 | 128 |
| Learning rate | 0.0008 | 0.001 |
| Optimizer | Adam | Adam |

Table 5: Best hyperparameters for the WGANs-GP trained on landmark genes.

| Hyper-parameter | GTEx | TCGA |
|---:|---:|---:|
| Batch size | 1024 | 64 |
| Layers disc. | [8192, 256] | [4096,128] |
| Layers gen. | [128, 1024, 4096] | [256, 1024, 8192] |
| Learning rate (disc.) | 0.000582 | 0.000124 |
| Learning rate (gen.) | 0.002205 | 0.000275 |
| Optimizer | RMSprop | RMSprop |
| Spectral normalization | False | False |

### 3. GMDA

Table 6: Best hyper-parameters of GMDA obtained after maximizing the precision-recall trade-off (F1 score) through a Bayesian search of 100 trials.

| Hyper-parameter | Diabetes | Gesture | Magic | Wilt | GTEx | TCGA |
|---|---|---|---|---|---|---|
| Layers | [32,64] | [256, 1024] | [256,256] | [128,128] | [128, 512, 2048] | [512, 1024, 1024] |
| Latent dim | 16 | 16 | 8 | 8 | 8 | 32 |
| Batch size | 400 | 1024 | 1024 | 2048 | 2048 | 2048 |
| Epochs | 2,500 | 2,000 | 1,500 | 1,500 | 5,000 | 5,000 |
| Learning rate | .0005 | .001 | .0005 | .005 | .001 | .0005 |
| Optimizer | RMSprop | Adam | RMSprop | RMSprop | RMSprop | Adam |
| $\delta$ (density) | .05 | .25 | .25 | .05 | .15 | .5 |
| $K$ (# probes) | 250 | 500 | 500 | 500 | 800 | 800 |
| $\eta$ (persistence) | .1 | .1 | .0 | .8 | .0 | .2 |
| $w_{DH}$ (dark probe) | 0. | 0. | 1. | 0. | 1. | 1. |

## C. Additional performance results

This section displays complementary details on the realism-diversity tradeoff (precision vs. recall) and the MLE performance using XGBoost.

Table 7: Precision results on real-life datasets (best results with statistical significance are highlighted in bold). GMDA ranks first, i.e. it generates the most faithful data w.r.t. the real distribution.

| Model | Diabetes | Gesture | Magic | Wilt | Avg. | Rank |
|---|---|---|---|---|---|---|
| TVAE | $\mathbf{99.39_{\pm 0.25}}$ | $24.08_{\pm 0.74}$ | $87.14_{\pm 0.19}$ | $94.97_{\pm 0.51}$ | 76.39 % | 3 |
| CTGAN | $0.73_{\pm 0.4}$ | $0.06_{\pm 0.06}$ | $20.07_{\pm 0.24}$ | $10.4_{\pm 0.61}$ | 7.82 % | 4 |
| TabDDPM | $65.95_{\pm 0.94}$ | $\mathbf{99.95_{\pm 0.01}}$ | $\mathbf{99.99_{\pm 0.01}}$ | $\mathbf{98.58_{\pm 0.22}}$ | 91.12 % | 2 |
| GMDA | $96.78_{\pm 0.78}$ | $84.61_{\pm 0.48}$ | $91.4_{\pm 0.18}$ | $97.52_{\pm 0.1}$ | **92.58%** | 1 |

Table 8: Recall results on real-life datasets (best results with statistical significance are highlighted in bold). It is important to note that this metric can be sensitive to spread outliers. This is why CTGAN ranks second while it performs the worst when considering precision (realism).

| Model | Diabetes | Gesture | Magic | Wilt | Avg. | Rank |
|---|---|---|---|---|---|---|
| TVAE | $91.69_{\pm 2.29}$ | $\mathbf{93.72}_{\pm \mathbf{0.42}}$ | $97.05_{\pm 0.11}$ | $97.47_{\pm 0.33}$ | **94.98%** | 1 |
| CTGAN | $77.96_{\pm 17.28}$ | $86.63_{\pm 1.}$ | $\mathbf{99.33}_{\pm \mathbf{0.14}}$ | $98.02_{\pm 0.68}$ | 90.49% | 2 |
| TabDDPM | $\mathbf{96.37}_{\pm \mathbf{0.7}}$ | $73.36_{\pm 0.49}$ | $93.71_{\pm 0.18}$ | $\mathbf{98.15}_{\pm \mathbf{0.22}}$ | 90.4% | 3 |
| GMDA | $89.04_{\pm 1.36}$ | $61.93_{\pm 0.92}$ | $90.57_{\pm 0.27}$ | $96.69_{\pm 0.21}$ | 84.56% | 4 |

Table 9: Machine Learning Efficiency (F1 score of XGBoost classifier) on real-life datasets (best results with statistical significance are highlighted in bold). GMDA ranks second overall, outperforming TabDDPM only on Wilt. On the other datasets, our method comes relatively close to the best F1 score.

| Model | Diabetes | Gesture | Magic | Wilt | Avg. | Rank |
|---|---|---|---|---|---|---|
| Baseline | $69.02_{\pm 1.21}$ | $42.96_{\pm 1.26}$ | $86.78_{\pm 0.41}$ | $91.74_{\pm 0.71}$ | - | - |
| TVAE | $65.79_{\pm 2.36}$ | $27.14_{\pm 1.42}$ | $81.94_{\pm 0.43}$ | $89.22_{\pm 1.37}$ | 66.02% | 3 |
| CTGAN | $43.07_{\pm 3.48}$ | $8.22_{\pm 0.0}$ | $52.7_{\pm 1.61}$ | $47.16_{\pm 1.85}$ | 37.79% | 4 |
| TabDDPM | $\mathbf{75.24}_{\pm \mathbf{1.24}}$ | $\mathbf{41.44}_{\pm \mathbf{0.81}}$ | $\mathbf{85.82}_{\pm \mathbf{0.32}}$ | $88.48_{\pm 0.59}$ | **72.74%** | 1 |
| GMDA | $67.58_{\pm 2.1}$ | $36.1_{\pm 0.84}$ | $84.48_{\pm 0.37}$ | $\mathbf{91.29}_{\pm \mathbf{0.64}}$ | 69.86% | 2 |

## D. Experimental results: PCA analysis

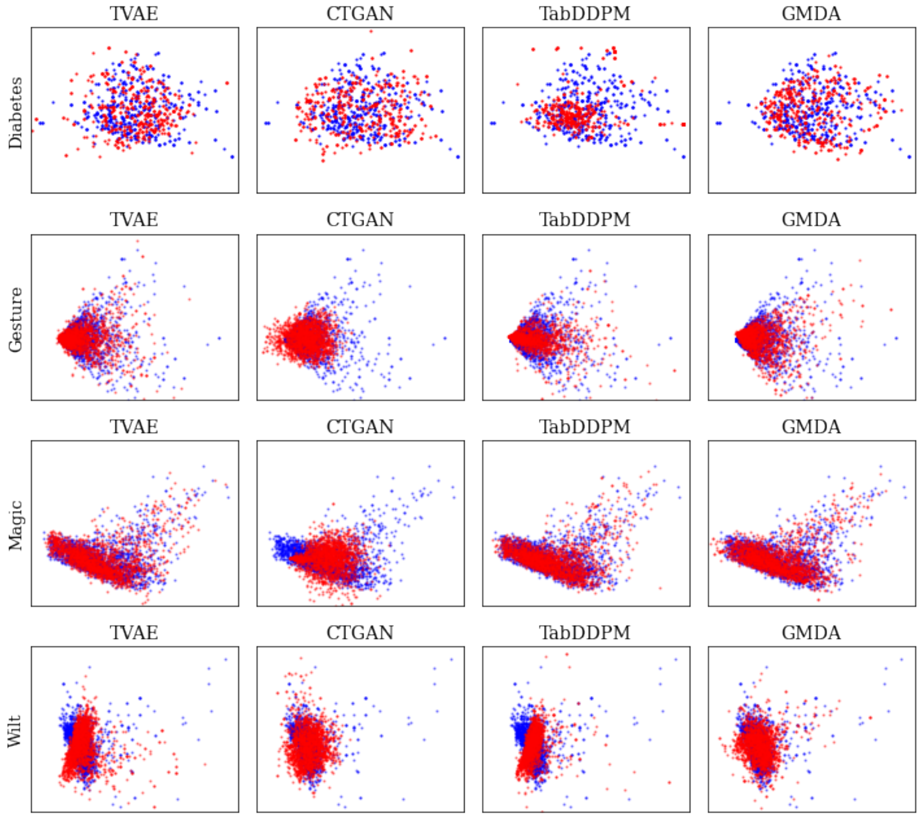


Fig. 1: PCA projections of real (blue) and generated (red) real-life datasets. We can see that CTGAN (on Gesture), TVAE (on Diabetes and Wilt), and TabDDPM (on Wilt) tend to generated out-of-distribution samples while GMDA-generated data holds similar structure in terms of principal components.

## E. Sensitivity analysis of the hyper-parameters

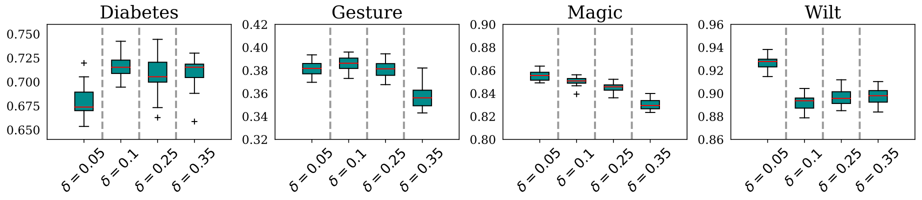
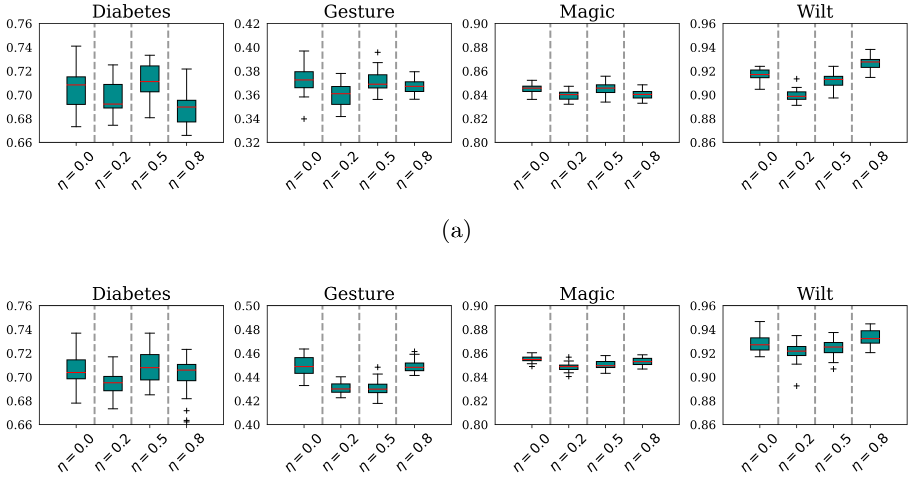


Fig. 2: GMDA: Sensitivity analysis of MLE w.r.t. the density $\delta$ parameter (XG-Boost F1 score, quantiles on 5 runs). Except for Diabetes, the other datasets benefit from smaller values of $\delta$ (.05).



(a)

(b)

Fig. 3: GMDA: Sensitivity analysis of MLE w.r.t. to the $\eta$ parameter (persistence): classification F1 score on test data for XGBoost (top) and Catboost (bottom). Depending on the dataset, no significant variation of performance is observed w.r.t. $\eta$. The fully stochastic regime ($\eta = 0.$) yields similar results compared to other values of persistence.

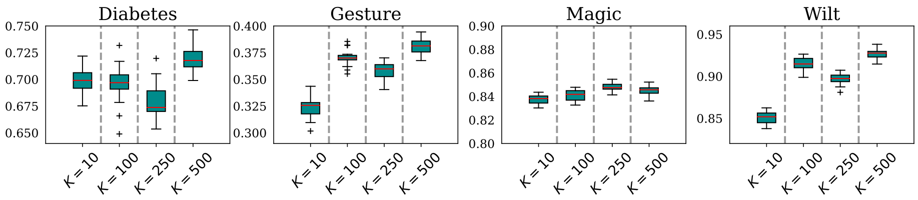
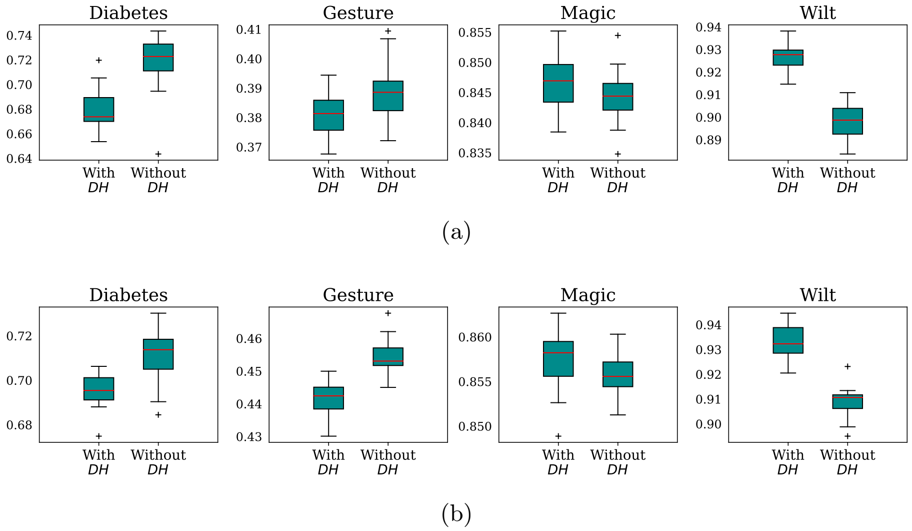


Fig. 4: GMDA: Sensitivity analysis of MLE w.r.t. the number $K$ of probes (XGBoost F1 score, quantiles on 5 runs). Except for Diabetes, the other datasets benefit from smaller values of $\delta$ (.05).



(a)

(b)

Fig. 5: GMDA: Sensitivity analysis of MLE w.r.t. the dark probe parameter: classification F1 score on test data for XGBoost (top) and for Catboost (bottom). The importance of the dark probe appears specific to each dataset as GMDA performs best without it ($w_{DH} = 0$) on Diabetes and Gesture, and better with it ($w_{DH} = 1$) on Magic and Wilt.

## F. Best settings classifiers

The best hyper-parameters for both XGBoost and Catboost classifiers were obtained after a Bayesian search of 20 runs to maximize the validation accuracy. For each of the high-dimensional gene expression datasets, a multilayer perceptron (MLP) was optimized through a Bayesian search of 1,000 trials over the validation set. These settings are the same ones used for training the classifiers on the synthetic data. The detail of the hyper-parameters can be found below

### 1. XGBoost

Table 10: Best hyper-parameters for XGBoost.

| Hyper-parameter | Diabetes | Gesture | Magic | Wilt |
|---|---|---|---|---|
| Gamma | 0 | 0 | 1 | 0 |
| Max depth | 10 | 10 | 10 | 10 |
| Min child weight | 10 | 20 | 5 | 5 |
| Number of estimators | 10 | 50 | 100 | 100 |

### 2. Catboost

Table 11: Best hyper-parameters for Catboost.

| Hyper-parameter | Diabetes | Gesture | Magic | Wilt |
|---|---|---|---|---|
| Leaf estimation iterations | 4 | 9 | 10 | 9 |
| Max depth | 3 | 8 | 8 | 7 |
| Learning rate | .23 | .03 | .04 | .39 |
| Bagging temperature | .65 | .06 | .13 | .77 |
| L2 leaf reg | 8.48 | 5.8 | 3.99 | 4.1 |
| Iterations | 2,000 | 2,000 | 2,000 | 2,000 |

### 3. MLP

Table 12: Best hyperparameters for the baseline MLP trained on high-dimensional datasets (landmark genes).

| Hyper-parameter | GTEx | TCGA |
|---|---|---|
| Batch size | 64 | 2048 |
| Dropout | 0.3 | 0.5 |
| Layers | [1024,1024] | [256,16] |
| Learning rate | 0.0001 | 0.004 |
| Optimizer | Adam | Adam |