Evaluation of Synthetic EHRs: Cystic Fibrosis Patients

Jer Hayes
Research Scientist, AI Labs Accenture, Dublin.

Objective

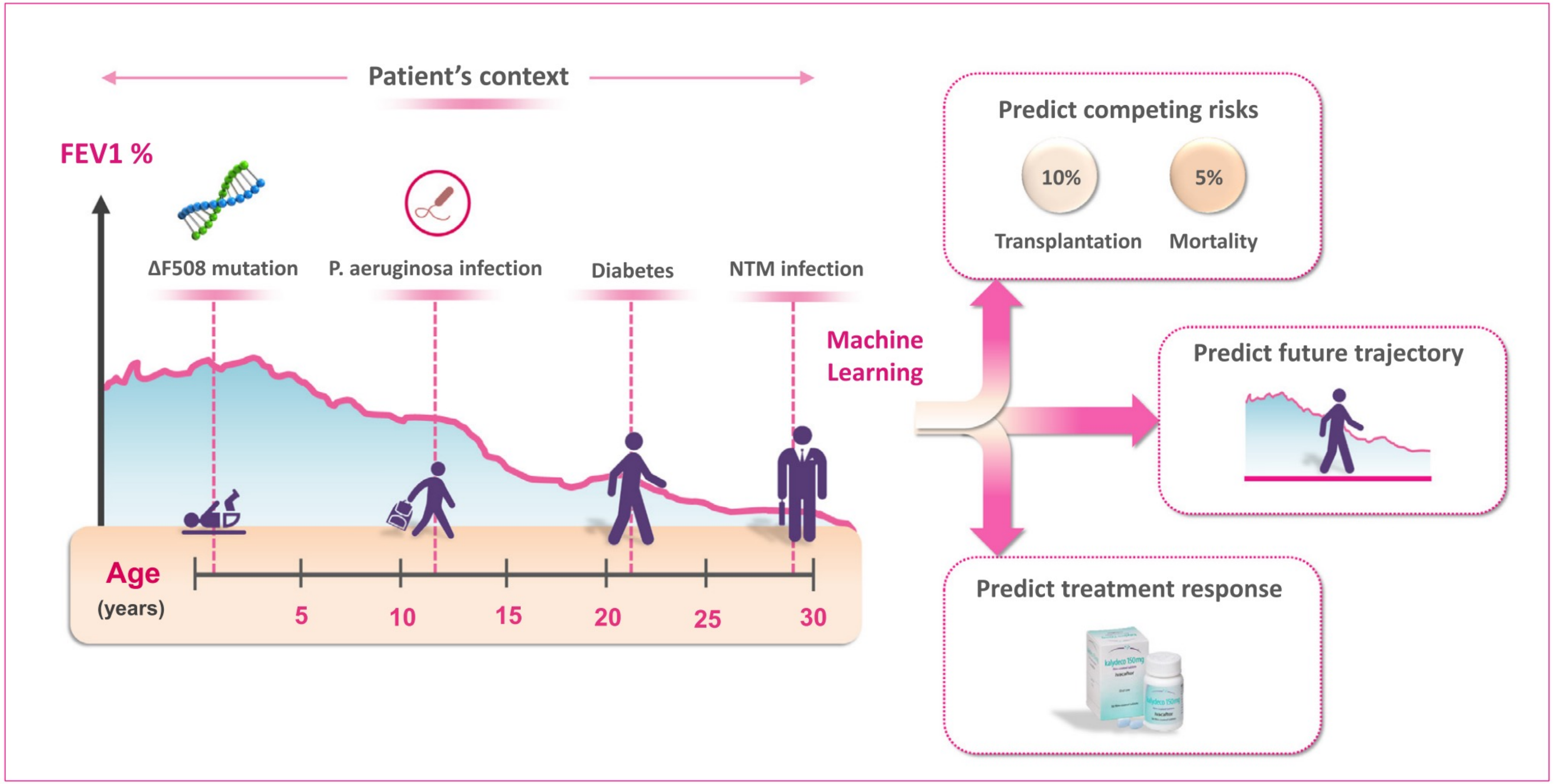
Goal:

- Can we create a ***faithful*** synthetic dataset for Cystic Fibrosis patients?

Motivation:

- Privacy concerns on sharing medical datasets.
- Data limitation and Data imbalance issues for Cystic Fibrosis patients.

The Case for AI & Health for Cystic Fibrosis



Opportunities for machine learning to transform care for people with cystic fibrosis [Abroshan, 2020].

accenture

Variable	Alive & no LT n = 8781 (%)	Death/LT n = 1293 (%)	UKCFΔ	USCFΔ
Gender(%male)	4,925 (56.1)	622 (48.1%)		
~~Age (years)~~§	55.6% missing			
~~Height (cm)~~§	81.0% missing			
~~Weight (kg)~~§	80.0% missing			
~~BMI (kg/m2)~~§	81.0% missing			
Genotype				
Homozygous	Not EHR			
Heterozygous	Not EHR			
ΔF508	Not EHR			
G551D	Not EHR			
Class I	Not EHR			
Class II	Not EHR			
Class III	Not EHR			
Class IV	Not EHR			
Class V	Not EHR			
Class VI	Not EHR			
Spirometry §				
~~FEV1 (L)~~	99.3% missing			
~~FEV1%~~	97.6% missing			
~~Best FEV1 (L)~~	99.3% missing			
~~Best FEV1%~~	98.7% missing			
~~FEV1% (2017)~~	99.5% missing			
~~FEV1% (2016)~~	99.5% missing			
~~FEV1% (2015)~~	99.5% missing			
~~FEV1% (2014)~~	99.5% missing			
Lung Infections				
~~B. Cepacia~~[a]	7 (0.01)	2 (0.1)	(5.2)	(1.6)
P. Aeruginosa	686 (7.8)	172 (13.3)	(52.8)	(23.7)
MRSA	831 (9.5)	172 (13.3)	(-5.6)	(9.7)
Aspergillus	132 (1.5)	62 (4.8)	(10.5)	
NTM	80 (1.0)	6 (0.5)	(4.2)	(9.0)
H. Influenza	72 (0.8)	25 (1.9)	(4.1)	(9.1)
E. Coli	203 (2.3)	50 (3.9)	(-2.0)	
K. Pneumoniae	72 (0.8)	26 (2.0)	(-0.6)	
Gram-negative	0	0	(0.5)	
ALCA	37 (0.4)	4 (0.3)	(2.3)	
Staph. Aureus	323 (3.7)	89 (6.9)	(19.9)	(52.9)
~~Xanthomonas~~[b]	60 (0.7)	9 (0.7)	(3.2)	
B. Multivorans	This is a UK concept			
B. Cenocepacia	This is a UK concept			
Pandoravirus	Not found			
Comorbidities				
Respiratory				
ABPA	138 (1.6)	17 (1.3)	(10.8)	(3.4)
Nasal Polyps	269 (3.1)	35 (2.7)	(0.0)	(6.8)
Asthma	1604 (18.3)	151 (11.7)	(-2.0)	(13.4)
Sinus Disease	750 (8.5)	110 (8.5)	(4.5)	(1.4)
Hemoptysis	398 (4.5)	85 (6.6)	(-3.2)	(-2.8)

Variable	Alive & no LT n = 8781 (%)	Death/LT n = 1293 (%)	UKCFΔ	USCFΔ
Pancreatic				
Cirrhosis	125 (1.4)	32 (2.5)	(1.2)	(1.6)
Liver Disease	309 (3.5)	68 (5.3)	(12.5)	(-0.1)
~~Pancreatitis~~[c]	11 (0.1)	2 (0.2)	(1.3)	(0.9)
~~Liver Enzymes~~[d]	5 (0.05)	0 (0)	(15.1)	
Gall Bladder	96 (1.1)	17 (1.3)	(-0.6)	
GI Bleed (variceal)	Not found			
Gastrointestinal				
GERD	1675 (19.1)	269 (20.8)	(1.6)	(17.4)
GIB (no variceal)	168 (2.0)	43 (3.3)	(-2.0)	
Intestinal Obstruction	358 (4.1)	125 (9.7)	(3.6)	
Musculoskeletal				
Arthropathy	440 (5.0)	62 (4.8)	(4.6)	(-1.8)
Bone Fracture	131 (1.5)	24 (1.9)	(-0.4)	(-1.3)
Osteopenia	4 (0.05)	1 (0.1)	(20.6)	(9.9)
Other				
Cancer	17 (0.2)	6 (0.5)	(0.1)	(-0.1)
Diabetes	1592 (18.1)	245 (18.9)	(9.0)	
CFRD	789 (9.0)	153 (11.8)	(23.2)	(9.3)
Pulmonary Abscess	32 (0.4)	15 (1.2)	(-0.4)	
Chr. Pseudomonas	549 (6.3)	148 (11.4)	(49.5)	
Osteoporosis	497 (5.7)	129 (10.0)	(3.4)	(-2.1)
AICU	Not found			
Kidney Stones	439 (5.0)	79 (6.0)	(-3.6)	(-4.4)
Cough Fracture	53 (0.6)	10 (5.5)	(-0.5)	
Hypertension	1784 (20.3)	285 (0.22)	(-14.9)	(-14.6)
A.Mycobacteria	132 (1.5)	18 (1.4)	(2.0)	(8.5)
Hearing Loss	255 (2.9)	59 (4.6)	(-0.4)	(-0.5)
Depression	1203 (13.7)	161 (12.5)	(-5.8)	(3.4)
Inhaled Antibiotics	Not found			
Muco-active Therapy				
DNase	252 (2.9)	17 (1.3)	(55.2)	(88.8)
Hypertonic Saline	1038 (11.8)	209 (16.1)	(11.6)	(61.3)
~~Promixin~~[e]	9 (0.1)	2 (0.2)	(20.5)	
Tobramycin	365 (4.2)	70 (5.4)	(-0.9)	(61.0)
iBuprofen	433 (5.0)	32 (2.5)	(-4.5)	(-3.8)
Oral Corticosteroids	Not found			
IV Antibiotics	Not found			
IV Antibiotic Courses				
Days at Home	Not found			
Days at Hospital	Not found			
Non-IV Hospitalization	Not found			
Non-IV Ventilation	Not found			
Oxygen Therapy	Not found			
Continuous	Not found			
Nocturnal	Not found			
Exacerbation	Not found			
Pro re nata	Not found			

A total of 10074 patients are extracted from the IBM Explorys database.

accenture

Data Processing

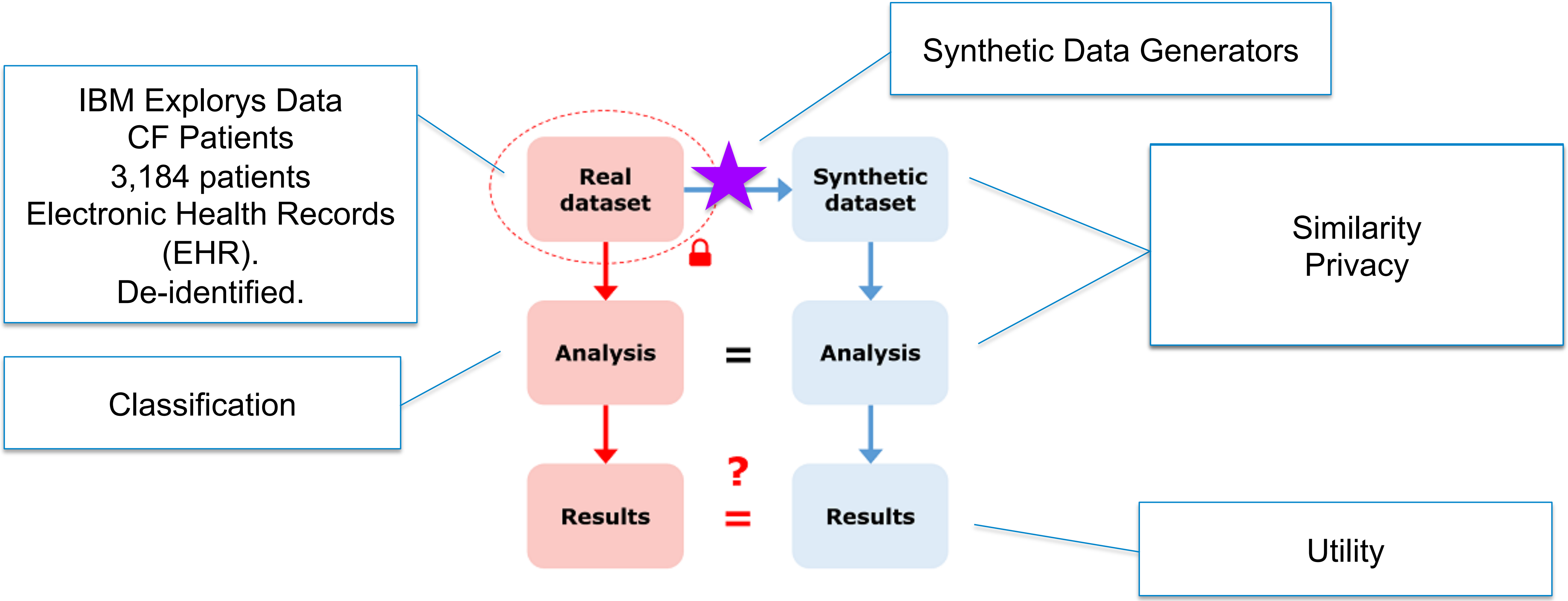
- Patients belong to two subgroups:
  - having died or having received a lung transplant, labelled by value 0;
  - having survived, labelled by value 1.

- We remove all samples with no related diagnosis codes and duplicates to enhance synthetic diversity.

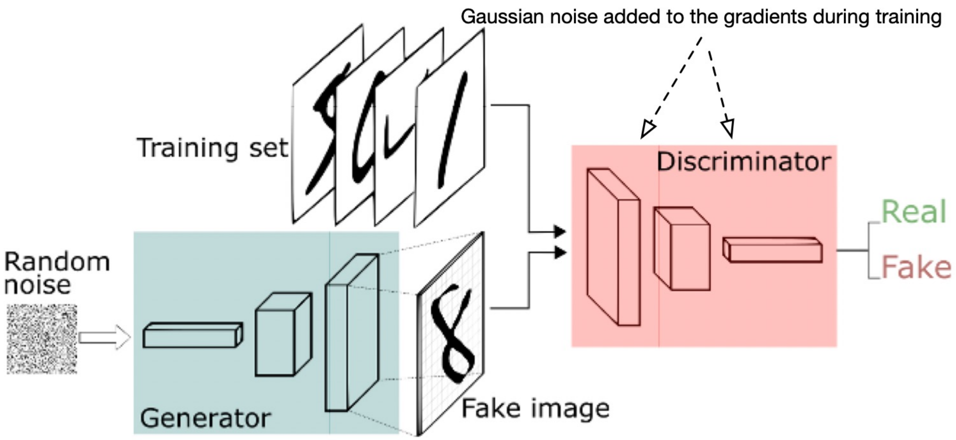
- For each patient, we assign value 1 to the diagnosis codes that have appeared in the medical history, and value 0 to these that have never appeared, resulting in a binary matrix of 41 variables.

- Our final dataset has 3184 patients, with ~ 80% belonging to the survived subgroup.

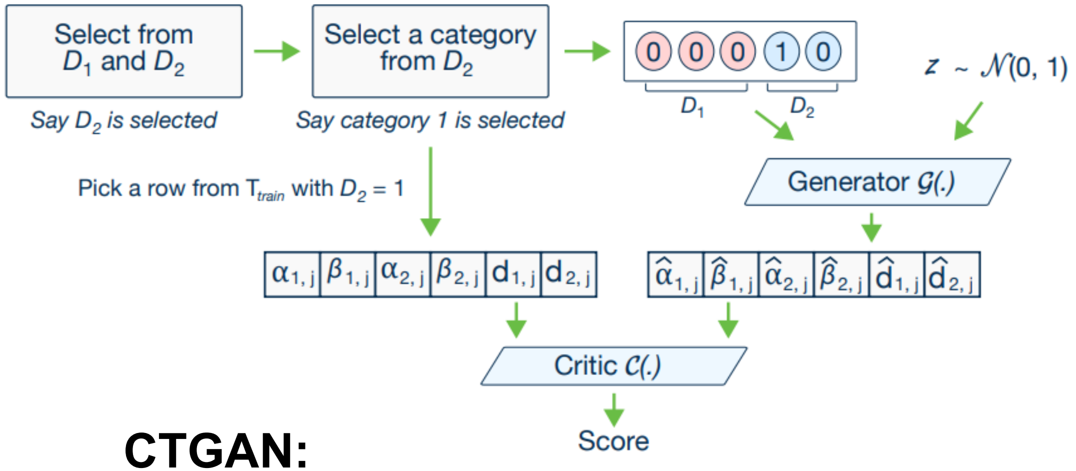
Synthetic Data Generation



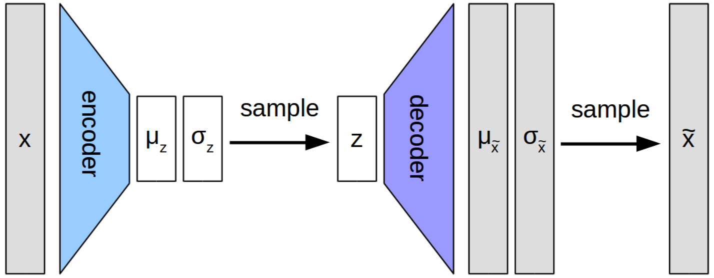
Synthesisers



DPGAN: Noise added to gradient during training



CTGAN:
Mixed categorical and cts features
Non-gaussian distributions/multi-modal
Sparse vectors
Imbalance in categorical features



VAE

accenture

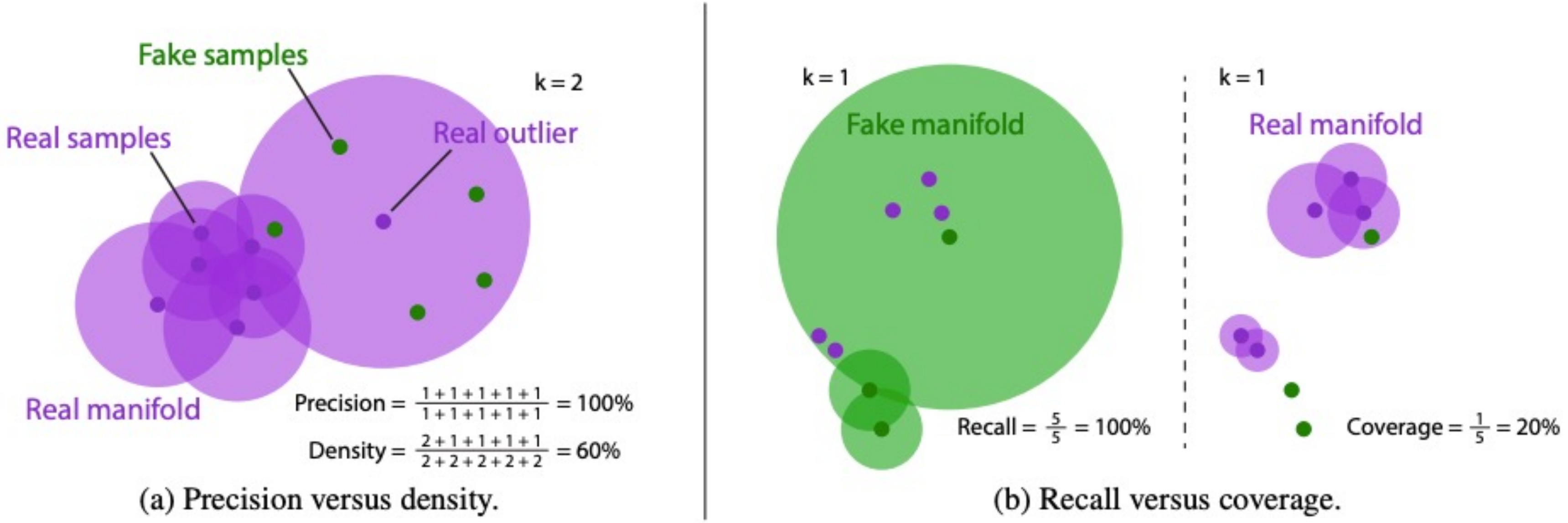
Faithful Evaluation

- **Similarity:** precision, recall, density and coverage
  - **Precision for fidelity** shows the degree to which generated samples resemble the real ones.
  - **Recall for diversity** means whether generated samples cover full variability of real ones.
  - **Density** rewards samples in regions where real samples are densely packed, relaxing the vulnerability to outliers.
  - **Coverage** improves upon the recall metric to better quantify this by building the nearest neighbour manifolds around the real samples, instead of the fake samples.

- **Uniqueness:** We consider the requirement of privacy as Uniqueness to not simply copy the input data.

- **Utility:** To empirically validate the Utility of the generated dataset, we evaluated the predictive ability of machine learning models trained on the synthetic datasets.

Naeem, Muhammad Ferjad, et al. "Reliable fidelity and diversity metrics for generative models." *International Conference on Machine Learning*. PMLR, 2020.



(a) Precision versus density.
(b) Recall versus coverage.

[Reliable Fidelity and Diversity Metrics for Generative Models (mlr.press)]()

accenture

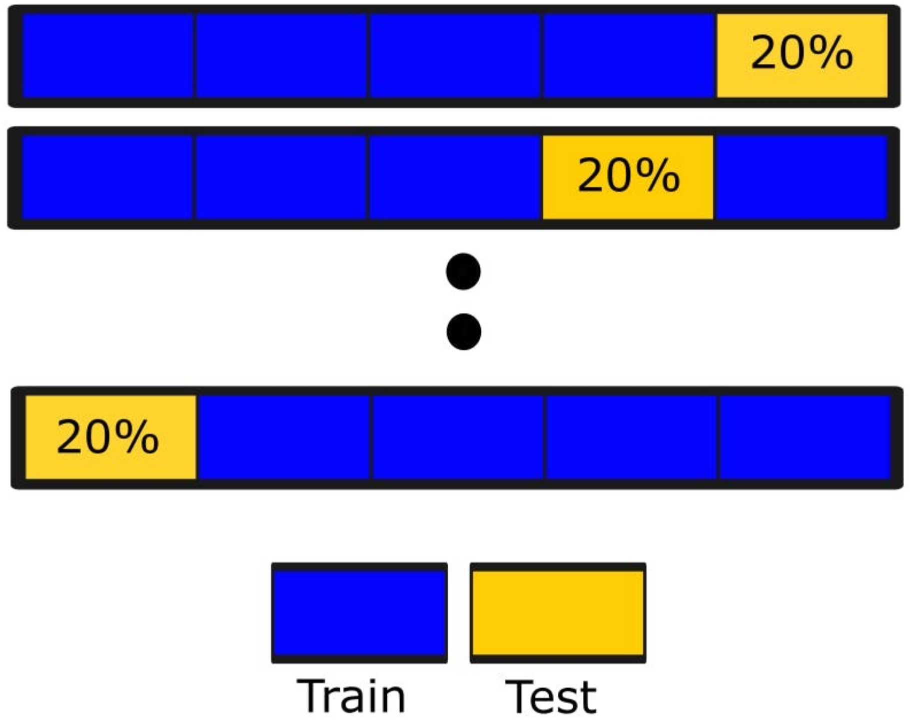
Experiments

Synthesis
Optimise hyperparameters using grid search with optimisation objective to maximise **Similarity** on random 80% stratified sample.

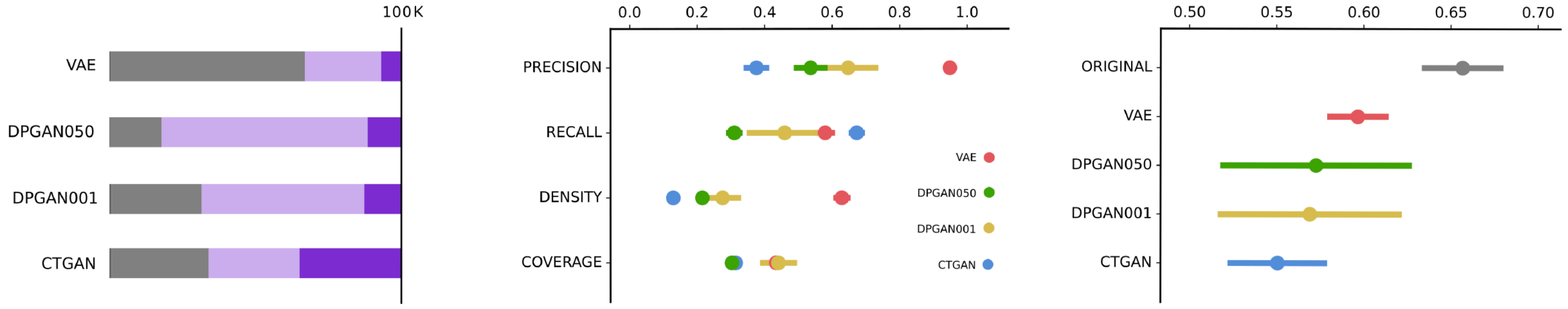
Setting A: Synthetic Dataset Only
Train classification models on synthetic training set, test the performance of the models on the real testing set.

Setting B: Synthetically Augmented Dataset
Augment class imbalance with synthesised records and repeat A.

5-fold cross validation stratified by outcome.



Similarity, Uniqueness and Setting A



(a) Authenticity proportion (b) Similarity metrics (c) Synthetic data AUC-ROC

Figure 1: (A) Authenticity of 100k samples from each generator. Grey is the proportion of samples that appear in the original training data. Light purple is samples that do not appear in the original training data, and darker purple represents those that are unique. (B) Similarity metrics for each model. For each fold, a dataset matching the size of the original fold with the equivalent proportion of classes is sampled from the unique synthetic dataset (dark blue only). This is repeated 10 times, and similarity metrics show mean and standard deviation over folds and repetitions.

Setting B – Balanced Class

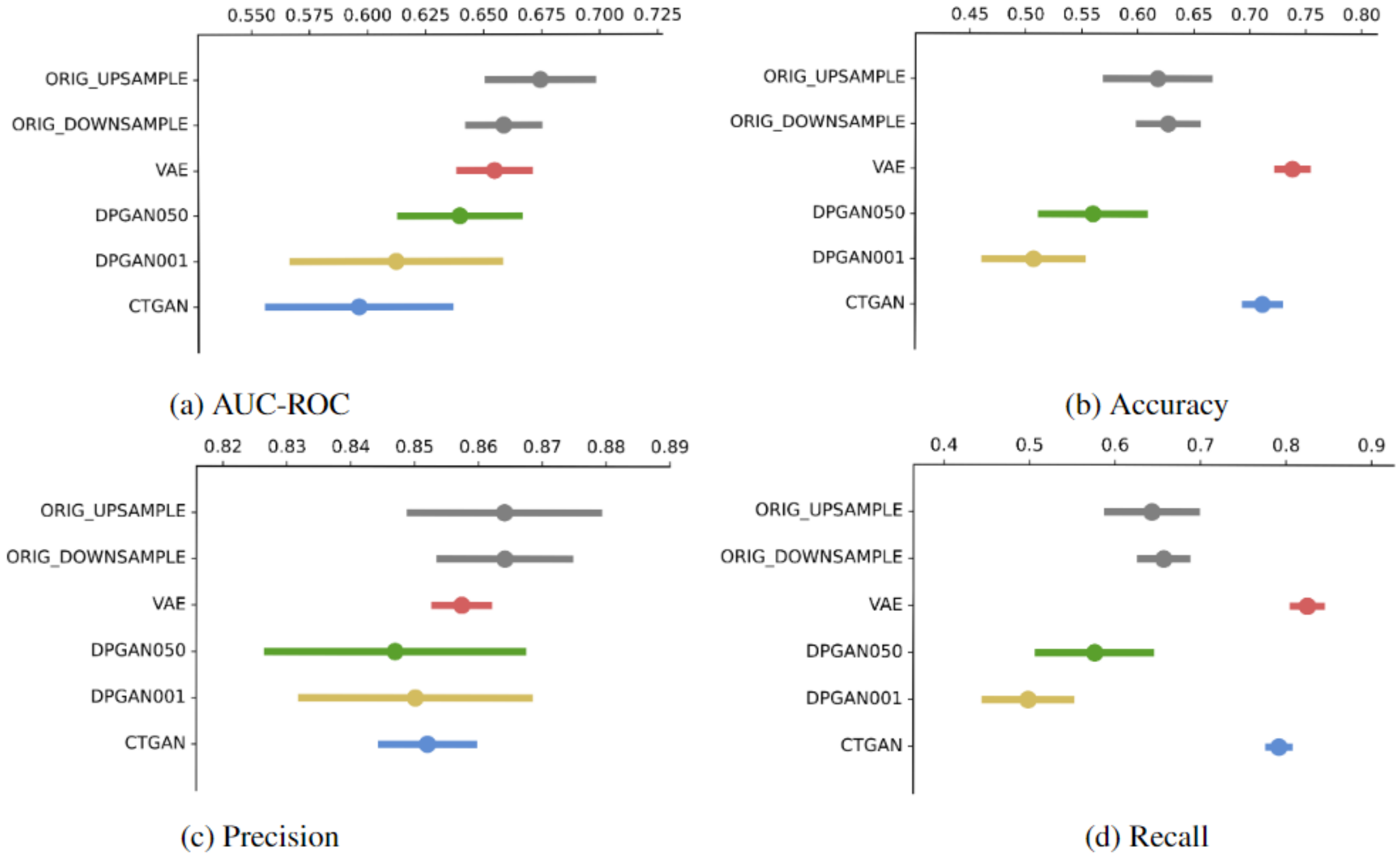
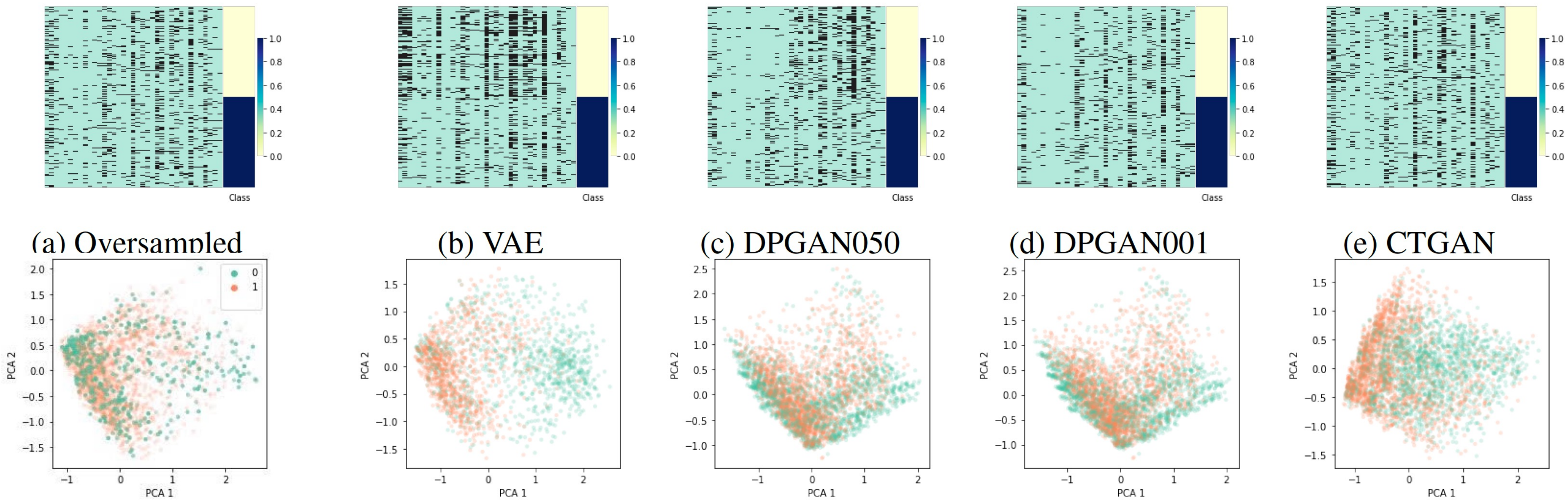


Figure 2: Synthetically Augmented Dataset

accenture

Visualising Similarity



(a) Oversampled (b) VAE (c) DPGAN050 (d) DPGAN001 (e) CTGAN

(f) Oversampled (g) VAE (h) DPGAN050 (i) DPGAN001 (j) CTGAN

Figure 3: The top row shows the heatmaps of binary matrices of sampled examples, where the columns represent different features and rows represent different samples. Class 0 at the top half of the matrix represents examples sampled from the dataset augmented with synthetic data, and Class 1 at the bottom half represents examples sampled from the original dataset.

Summary

- We observed increased accuracy in performance for both VAE and CTGAN.

- Diversity of CTGAN and VAE give rise to greater separability, resulting higher recall.

- CTGAN has highest uniqueness, which is beneficial for considering stricter conditions on privacy.

accenture

Thank you.

Jer Hayes -
jeremiah.hayes@accenture.com

accenture