GAN-Based Training for Binary Classifier

Mini Project

Outline

1. Challenge Statement
2. Implementation
3. Quality of learnt distribution
4. Further considerations
5. Model

Challenge Statement

The challenge is to use pictorial-based or raw data-based GAN to generate malign samples suitable for classifier training

Implementation (Trials)

- Classifiers
  - Linear Regression
  - Support Vector Machine (SVM)
  - Extra Trees
  - Random Forest
  - XGBoost

- GAN Models
  - CTGAN
  - CopulaGAN
  - Vanilla GAN (ReLU)
  - Vanilla GAN (LeakyReLU)
  - WGAN (LeakyReLU)
  - Pictorial GAN (CNNs)

Generators models performance

Median of Delta values:

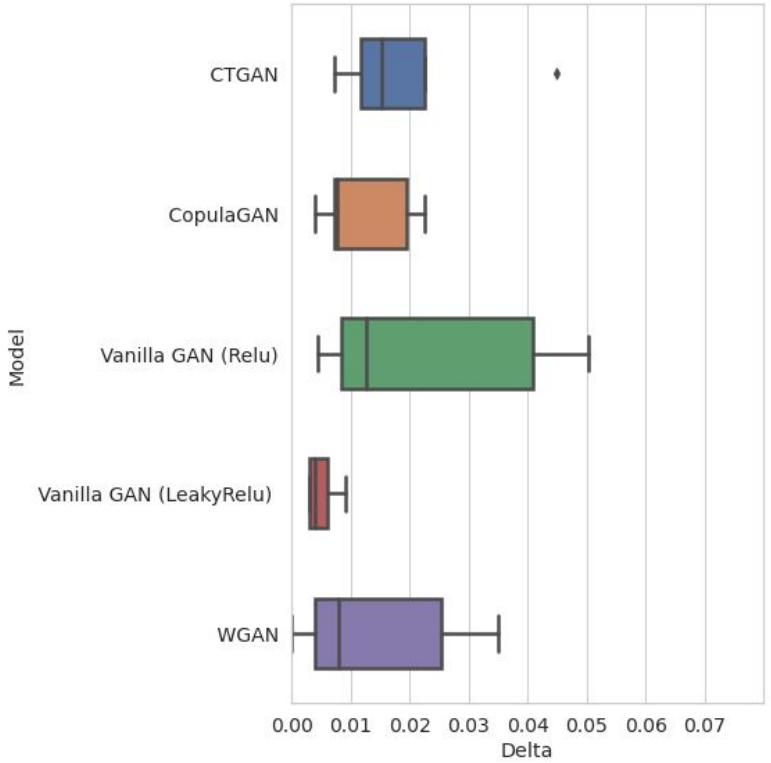
CTGAN = **0.0153**

COUPLAGAN = **0.0076**

Vanilla GAN (Relu) = **0.0125**

Vanilla GAN (LeakyRelu) = **0.004**

WGAN = **0.0078**



الخدمة الوطنية والإحتياطية

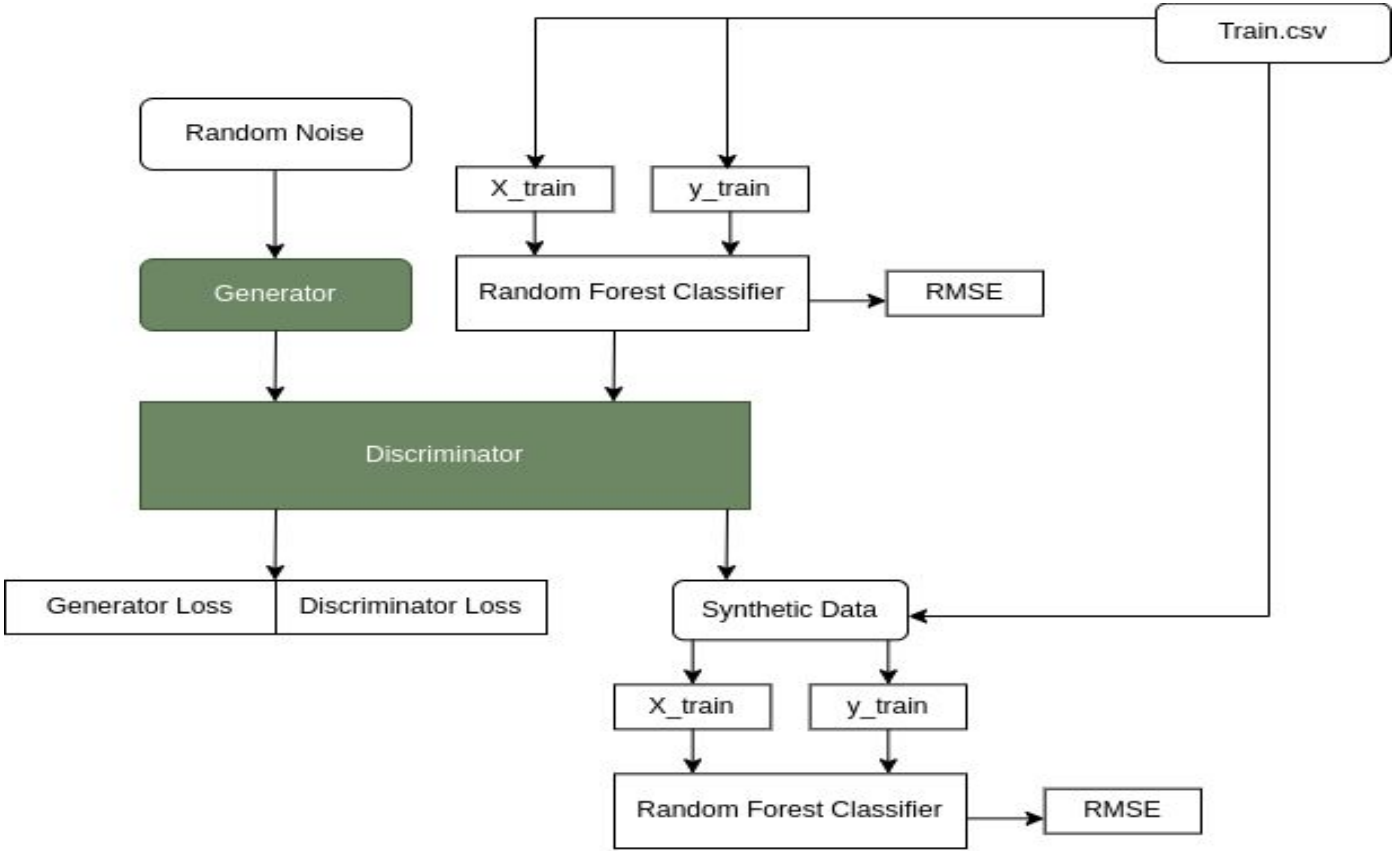
جامعة خليفة
Khalifa University

Extra Trees		
Before GAN	After GAN	Delta
12.978 %	11.698 %	0.013
10.26 %	11.239 %	0.01
12.978 %	10.761 %	0.022
10.26 %	11.698 %	0.14
11.239 %	11.239 %	0
		0.037

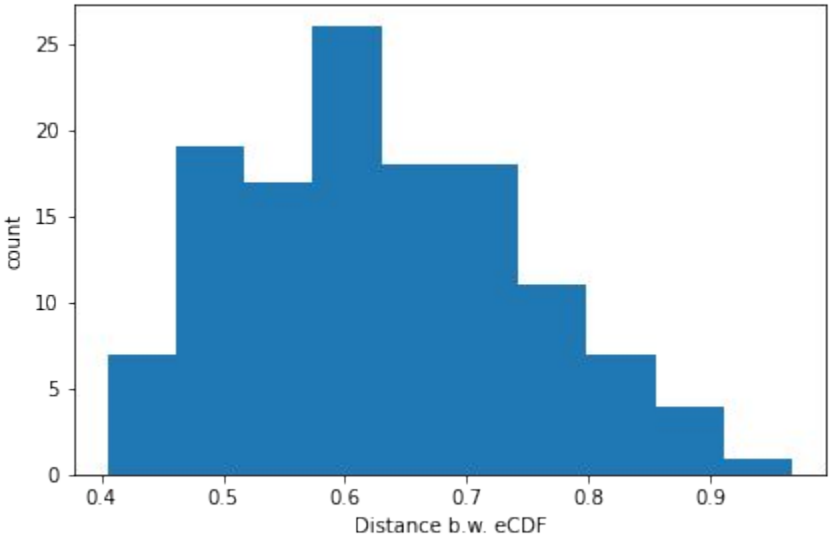
Random Forest		
Before GAN	After GAN	Delta
13.377 %	13.764 %	0.004
17.77 %	18.35 %	0.006
13.765%	13.765 %	0.00
16.859 %	17.168 %	0.003
16.859 %	16.543 %	0.003
		0.0032

XGBoost		
Before GAN	After GAN	Delta
18.377 %	12.566 %	0.008
18.377 %	12.978 %	0.004
18.377 %	10.761 %	0.026
18.377 %	13.377 %	0
18.377 %	11.239 %	0.021
		0.0118

Flow Chart


Train.csv
Random Noise
X_train
y_train
Generator
Random Forest Classifier
RMSE
Discriminator
Generator Loss
Discriminator Loss
Synthetic Data
X_train
y_train
Random Forest Classifier
RMSE

Learnt distribution: Kolmogorov-Smirnov



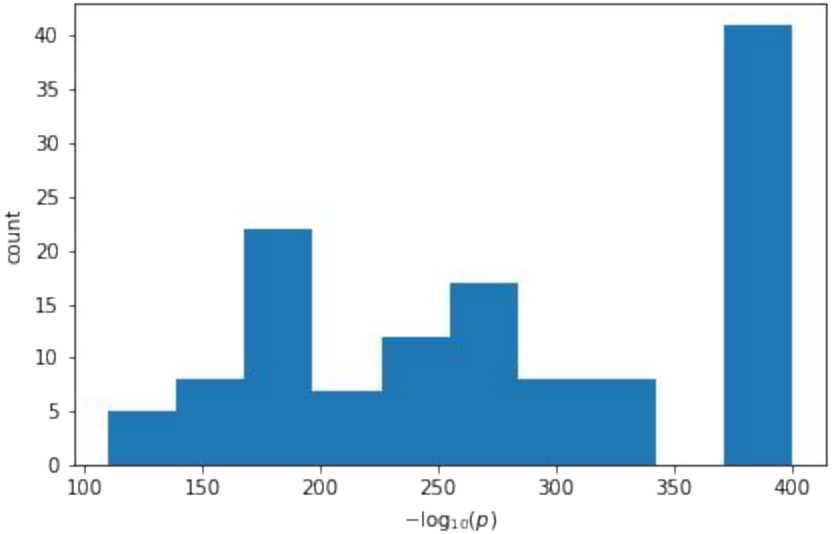
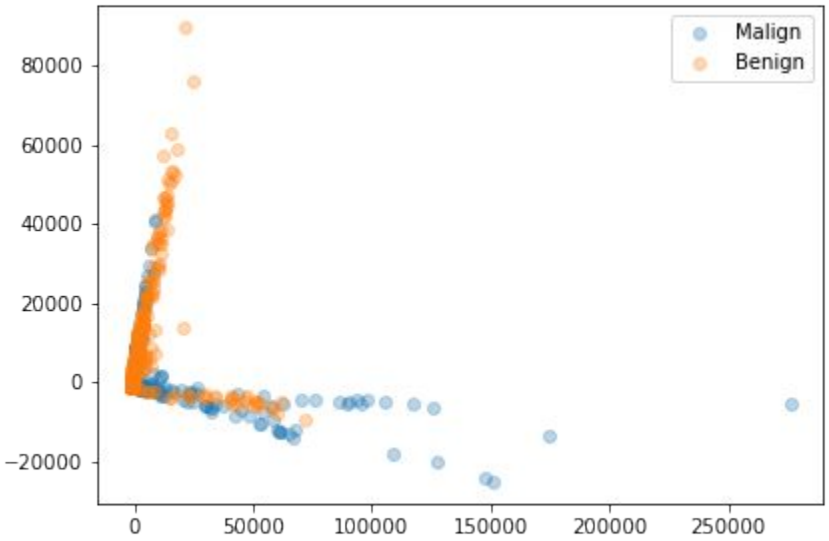


Fig 1: D-statistic i.e. the max distance between eCDFs
Fig 2: Statistical confidence in rejecting the null hypothesis

Learnt distribution: PCA



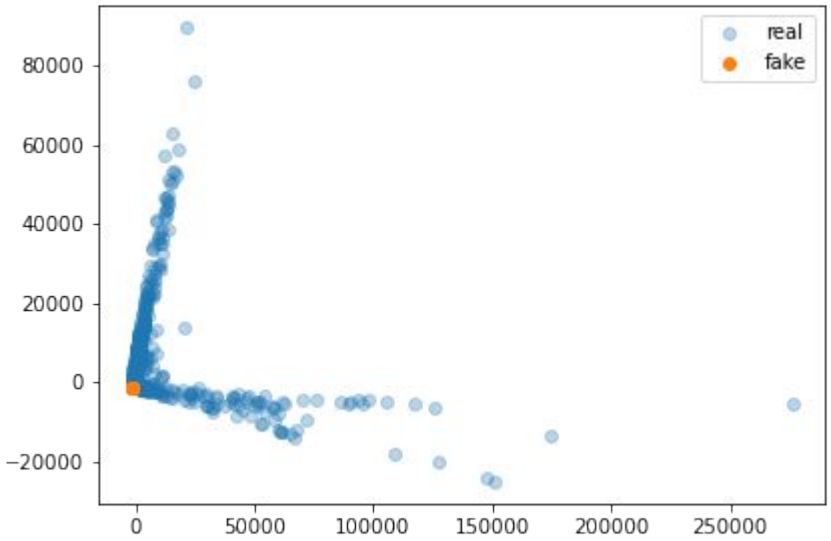


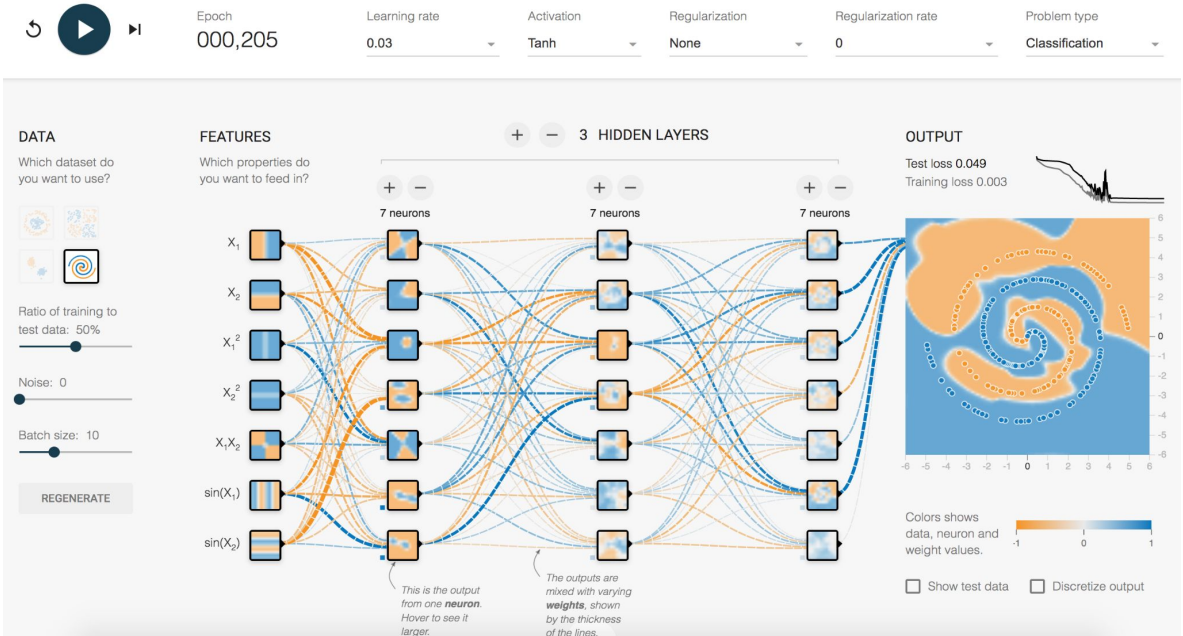
Fig 1: distribution of labels.
Fig 2: synthetic data projected onto PCA of real data.

LeakyReLU: **S**imple, **PAR**ametric **GAN** Model

1. Layer Features:
   a. Dense
   b. BatchNormalization
   c. Dropout

Some considerations

- Outlier Detection & Removal
- Feature interactions (e.g. x_1^2, x_2^2, x_1x_2, $sin(x_1)$, $sin(x_2)$)
- Explore other Activation Functions (e.g. Bionodal Root Unit (BRU))



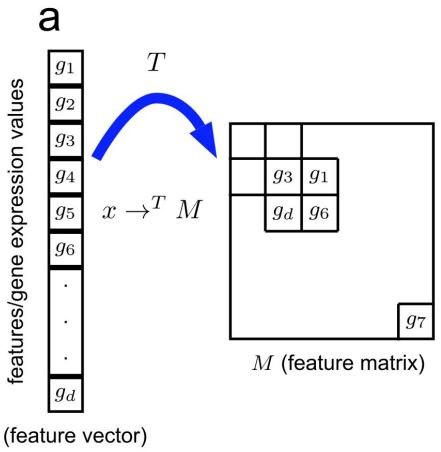
الخدمة الوطنية والإحتياطية

جامعة خليفة
Khalifa University

Implementation (Pictorial)

- Why consider pictorial? Why CNNs?
- How to transform tabular data to images
  - DeepInsight: kPCA/t-SNE

- From a 128 vector feature to 12x12x1 image
- Used conditional GAN to balance the dataset

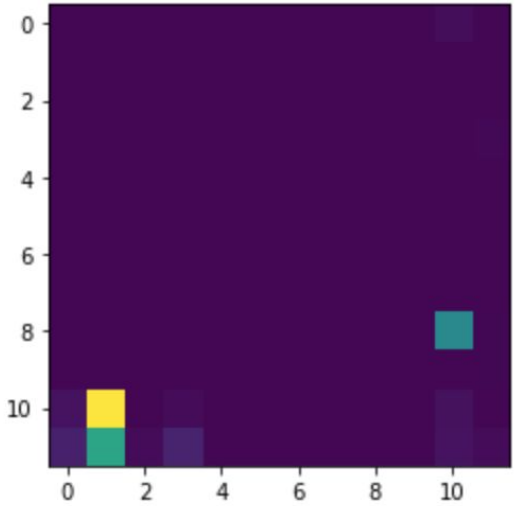


From feature vector x to a feature matrix M [1].

- Tough problem to fine-tune (number of pixels, 1 vs. 3 channels, etc)
- Tabular method achieves better performances (0.146)

[1] Sharma, Alok, et al. "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture." *Scientific reports* 9.1 (2019): 1-7.

Implementation (Pictorial)



Example of a malign synthetic image

الخدمة الوطنية والإحتياطية

جامعة خليفة
Khalifa University

The Team

Name	ID:
Abdelrahman Alblooshi	100062301
Hamad Alsheraifi	100062314
Saoud Sharif	100062324
Ali Alhashmi	100062327
Saeed Aljaberi	100062328
Hussain Sajwani	100062332


الخدمة الوطنية والإحتياطية

Questions?


جامعة خليفة
Khalifa University