جامعة خليفة
Khalifa University

IEEE Bigdata Cup 2022, Osaka, Japan

‘Using Generative Adversarial Networks to Simulate System Calls of Malicious Android Processes’

Challenge: ‘IEEE BigData Cup 2022: GAN-Based Training for Binary Classifier’
Paper ID: SC04203

Hamad Alsheraifi, Hussain Sajwani, Saeed Aljaberi, Abdelrahman Alblooshi, Ali Alhashmi, Saoud Sharif, Prof. Ernesto Damiani

100062314@ku.ac.ae, 100062332@ku.ac.ae, 100062328@ku.ac.ae, 100062301@ku.ac.ae, 100062327@ku.ac.ae, 100062324@ku.ac.ae, ernesto.damiani@ku.ac.ae

جامعة خليفة
Khalifa University

Outline

1. Introduction
2. Objective
3. Literature Review
4. Main Research
5. Learnt Distribution Analysis
   a. Kolmogorov-Smirnov
   b. PCA
6. Further Work

جامعة خليفة
Khalifa University

Introduction

- **Issue** with Binary Classification
  - Data Imbalances
  - End Result: Bias between majority class and minority class
- Vitally **Essential** to tackle this problem
- **End Goal**: Provide a Benchmark to determine Optimal Model for binary classification

جامعة خليفة
Khalifa University

Objective

- Gathering the training malware traces is restive and can be a nuisance depending on the type of malware.
  - Behavioral polymorphism
- Developing a GAN model to generate a proper training mechanism for binary classification.
- GANs are well suited for these issues because they can generate synthetic data that mimic actual data.

Key Words:-

Generative Adversarial Networks (GANs), Deep Learning, System Calls, Android Malware Detection

جامعة خليفة
Khalifa University

Literature Review

Context of the Problem:

- Cyber services
- Artificial Intelligence (AI) And Machine Learning (ML)

Why it Matters:

- Malicious techniques
- Detecting malicious processes

Proposed Solution:-

- Generating Synthetic Data using GAN

جامعة خليفة
Khalifa University

Main Research

Brief GAN Overview

- Deep learning algorithm can generate synthetic data with the same characteristics as the trained data
- Tabular-based and the Pictorial-based GAN models

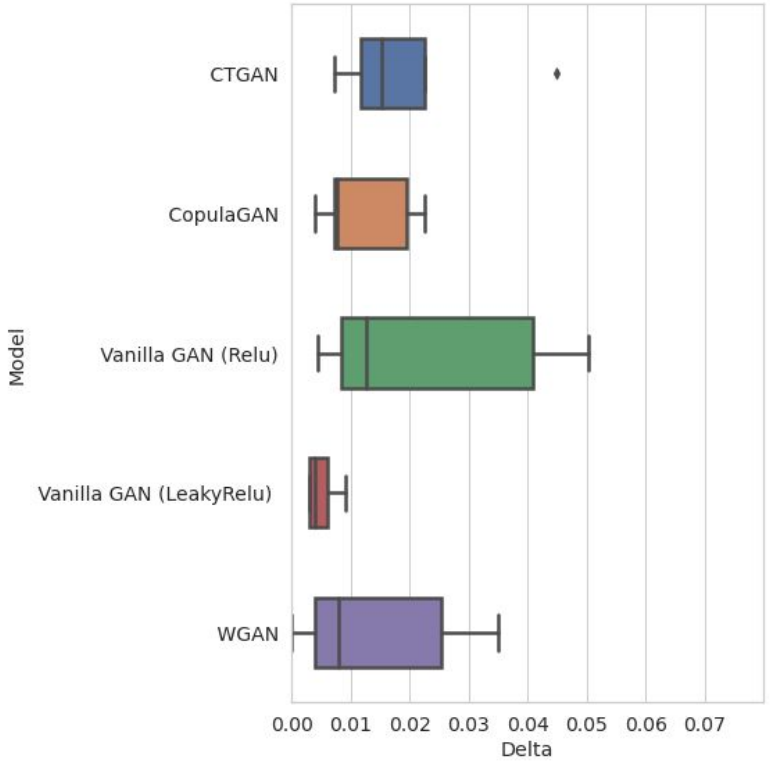
Classifiers	GAN Models
AdaBoost	CTGAN
Bagging	CopulaGAN
Extra-Trees	Vanilla GAN (ReLU)
Gradient Boosting	Vanilla GAN (Leaky-ReLU)
Random Forest	WGAN (Leaky-ReLU)
Stacking	Pictorial GAN (CNNs)
Voting	
Histogram-Based Gradient Boosting	
Support Vector Machines (SVMs)	
Linear Regression	

جامعة خليفة
Khalifa University

Main Research

Tabular GAN models (Median of delta values)

- CTGAN = (0.0153)
- CopulaGAN = (0.0076)
- Vanilla GAN (ReLU) = (0.0125)
- Vanilla GAN (LeakyReLU) = (0.004)
- WGAN (LeakyReLU)= (0.0078)



جامعة خليفة
Khalifa University

Results of Machine Learning Algorithms

Testing parameters:

- 100 epochs
- 10 runs

Random Forest			Extra Trees			XGBoost		
Before GAN	After GAN	Delta	Before GAN	After GAN	Delta	Before GAN	After GAN	Delta
13.377 %	13.764 %	0.004	12.978 %	11.698 %	0.013	18.377 %	12.566 %	0.008
17.77 %	18.35 %	0.006	10.26 %	11.239 %	0.01	18.377 %	12.978 %	0.004
13.76%	13.765 %	0	12.978 %	10.761 %	0.022	18.377 %	10.761 %	0.026
16.859 %	17.168 %	0.003	10.26 %	11.698 %	0.14	18.377 %	13.377 %	0
16.859 %	16.543 %	0.003	11.239 %	11.239 %	0	18.377 %	11.239 %	0.021
		0.0032			0.037			0.0118

Main Research, continued

Pictorial Representation

- Feature extraction is performed by the CNN's hidden layers
- Transforming feature vectors (or tabular data) is done by performing dimensionality reduction techniques, such as t-SNE or kPCA
- Hard problem to fine-tune (number of pixels,channels, etc)
- Tabular representation held better results

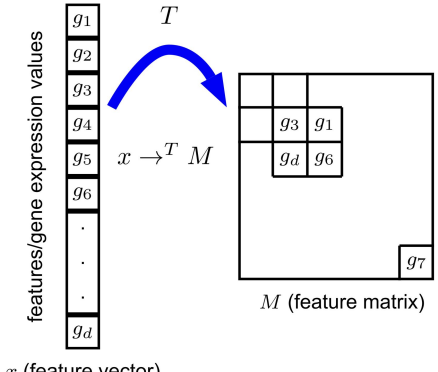
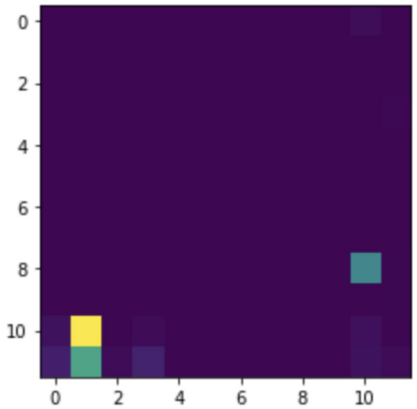


Figure taken from [1]

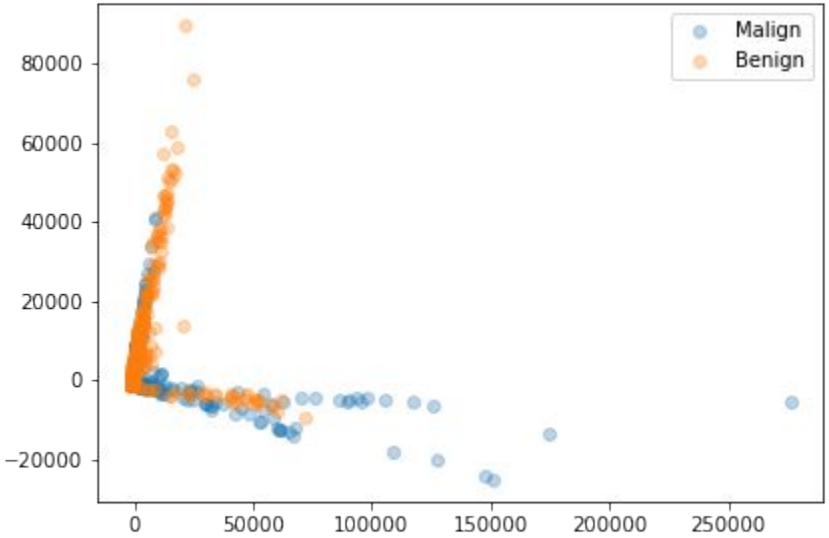


Example of malign synthetic image

[1]: Sharma, A., Vans, E., Shigemizu, D. *et al.* DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Sci Rep* 9, 11399 (2019).

جامعة خليفة
Khalifa University

Learnt malicious processes distribution: PCA



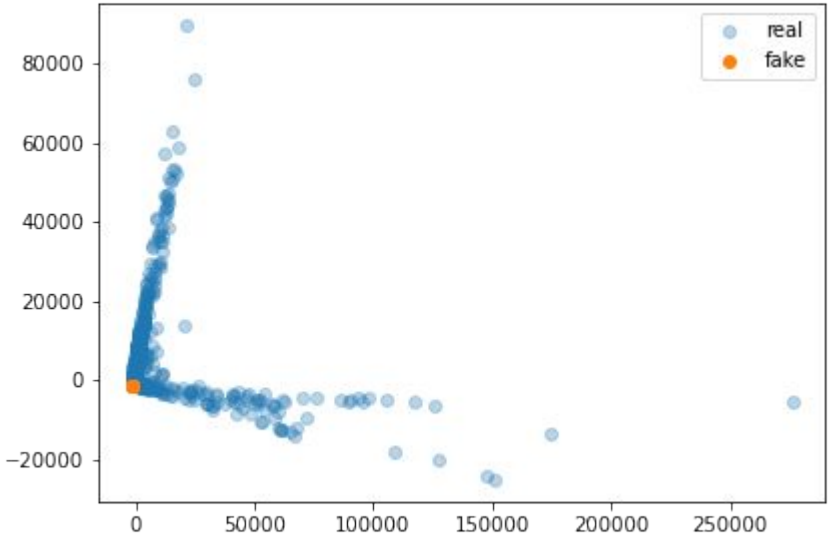
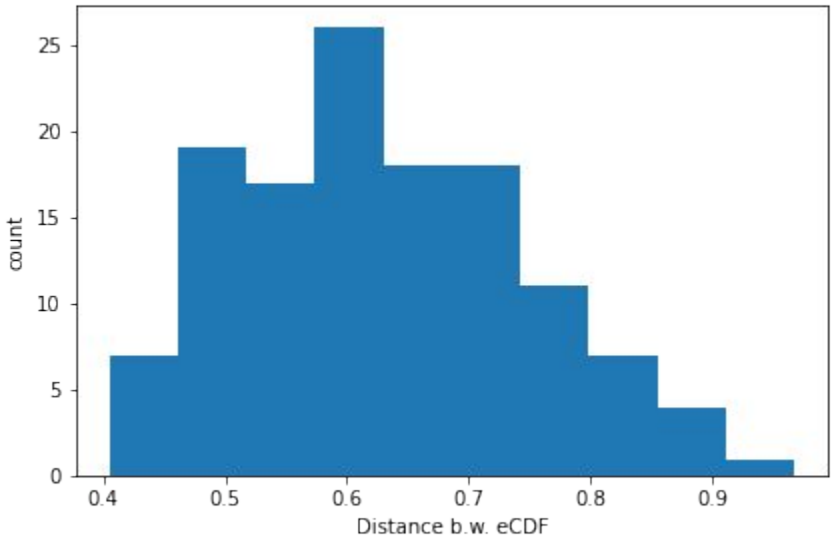


Fig 1: Distribution of labels.
Fig 2: Synthetic data projected onto PCA of real data.

جامعة خليفة
Khalifa University

Learnt malicious processes distribution: Kolmogorov-Smirnov



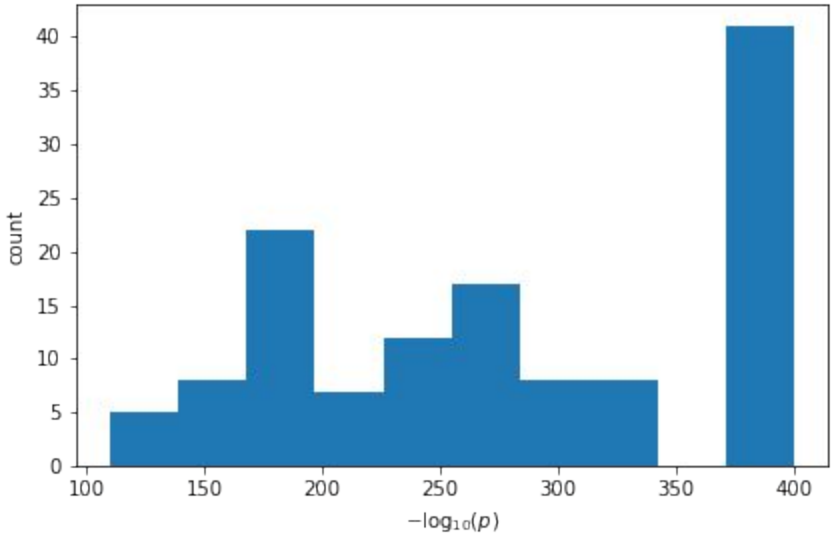


Fig 1: D-statistic i.e. the max distance between eCDFs
Fig 2: Statistical confidence in rejecting the null hypothesis

جامعة خليفة
Khalifa University

Further work

- Data and data representation:
  - The current data only uses monograms of system calls. Future work could include looking at longer n-grams.
  - This work only looks at generation of system calls. However, a more holistic approach would be including things like meta data of processes as well as API calls made by the processes.
- Model level optimization
  - Different GAN architectures
  - Further hyperparameter tuning

جامعة خليفة
Khalifa University

Thank you for your time