# A GAN-Assisted Data Quality Monitoring Approach for Out of Distribution Detection


**Kent Slater**

**Collaborators: Dr. Chenang Liu (Faculty Advisor), Yuxuan Li, and Dr. Yongwei Shan**

**Smart Technology & Analytics Research (STARs) Laboratory**

**School of Industrial Engineering and Management, Oklahoma State University**


## Research Objective

Develop an **anomaly detection approach** that tackles the challenges of high dimensionality, small sample size, unknown underlying distributions, and high levels of noise integrating the Generative Adversarial Network (GAN), a k-Nearest Neighbor (k-NN) based algorithm, and the Control Chart.

## Research Overview

- Data quality monitoring plays a critical role in various real-world product inspection problems
- Anomalous or invalid inspection data commonly exist due to recording errors, sensors, faults, etc.
- Approach utilizes the following to solve these challenges:
  - **GAN** learns underlying distribution and eliminates noise
  - **k-NN** measures similarity among data points
  - **Control Chart** provides out of distribution detection
- Approach is **scalable** to high dimensional data

## Background/Methodology

### What is the GAN?

- GAN consists of two neural networks, a ***Generator*** to generate data similar to real data, and a ***Discriminator*** to distinguish generated data from real data.
- The Generator and Discriminator compete against each other getting better at making generated data until it is indistinguishable from the real data.

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim P_{\text{data}}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{z}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

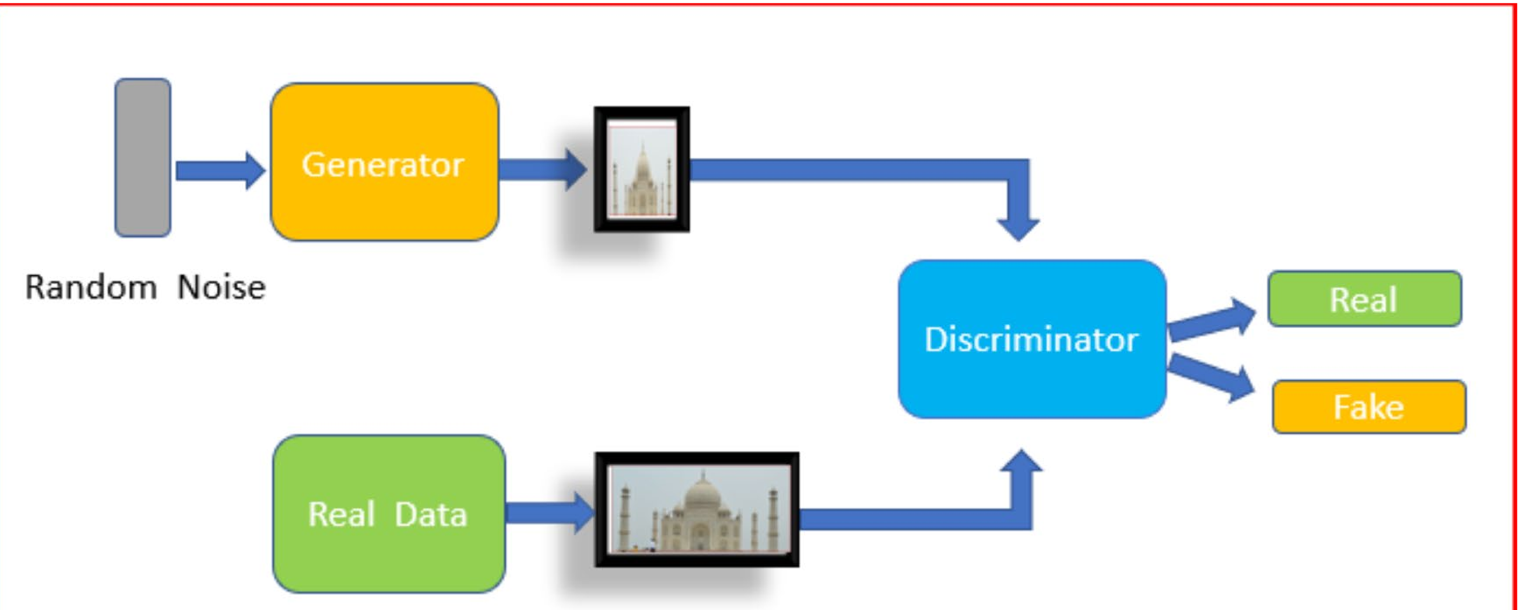


Figure 1: GAN Illustration

Source: https://www.analyticsvidhya.com/blog/2021/04/generate-your-own-dataset-using-gan/

### Why use the GAN?

- Raw data maybe be limited, **GAN can increase sample size**
- GAN can **eliminate noise**
- Therefore, accurately learned GAN-generated data will be **more effective for control chart development**

**Raw data** → **Train GAN** *(learn data distribution)* → **Generate Data** *(sufficient denoised data)* → *Control chart development*

Figure 2: GAN Training Flowchart

## Methodology

### K –Nearest Neighbor Algorithm

1. The Euclidian distances from each GAN-generated points are calculated
2. The k-Nearest Neighbors are determined
3. The average distance from the point to its k-nearest neighbors is calculated, $\bar{d}_i$
4. The grand average of $\bar{d}_i$ is calculated, $\bar{\bar{d}}$

### Control Chart Construction

- Control limits can be defined through quantiles or the Shewhart model
- For the quantile-based control limit, it could be defined using the upper 95th and 99th percentile of the $\{\bar{d}_i, i = 1,2,\dots,M\}$
- If the $\bar{d}_i$ of a new coming point is greater than the specified quantile, then it would be identified as out of control, i.e., out of distribution
- For the Shewhart model, the control limits can be defined as follows where $\hat{\sigma}_d$ represents the estimated standard deviation for $\{\bar{d}_i, i = 1,2,\dots,M\}$

$$\text{UCL} = \bar{\bar{d}} + k\hat{\sigma}_d$$
$$\text{LCL} = \bar{\bar{d}} - k\hat{\sigma}_d$$

- $k$ is the distance of the control limits from the center line, expressed in the units of $\hat{\sigma}_d$

**Quality Characteristic**
- *Identify k-nearest neighbors*
- *Calculate average distance $\bar{d}_i$*

→ **Center Line** *Overall average $\bar{\bar{d}}$*

→ **Control Limits**
- *Upper quantile, or*
- *Shewhart model*

→ *Out of distribution detection*

Figure 3: Control Chart Construction Flow Chart

## Experimental Setup

- 3 numerical experiments (Bivariate, Multivariate, and Gaussian process with window size 10), 10 trials each and $k = 3$
- 1 Training set, and **2 Validation** sets with 1000 points each
  - **One** with same distribution "normal data"
  - **One** with data with a slightly different distribution "abnormal data"

### Case 2 (Multivariate (1000,5))

| Training and Normal Data | Anomalous Data |
|---|---|
| $x_1 \sim U(0,1)$<br>$x_2 = 3x_1 + \epsilon, x_3 = x_2/3 - \epsilon$<br>$x_4 = 3x_3 - \epsilon, x_5 = x_4/3 - \epsilon$ | $x_1 \sim U(0,1)$<br>$x_2 = 5x_1 - \epsilon, x_3 = x_2/5 + \epsilon$<br>$x_4 = 5x_3 - \epsilon, x_5 = x_4/5 + \epsilon$ |

### Case 3 (Gaussian Process, (1000,10))

- The training data and normal data were sampled from the distribution:

$$\mathbf{x} \sim \text{GP}(0,\kappa),\ \kappa(x_{l_1}, x_{l_2}) = \exp\left(-\frac{1}{2\theta}\left(\left\|x_{l_1} - x_{l_2}\right\|_2^2\right)\right)$$

- The simulation adopted the radial basis function (RBF) kernel, and the parameter $\theta$ in the RBF kernel is $10^{-10}$. $l_1, l_2 = 1,2,\dots,10000$.
- The anomalous data were sampled from same distribution but with parameter $\theta$ equal to $2 \times 10^{-6}$ and shifted upward 1.5 units

## Results

### Performance of GAN to Learn Data Distribution

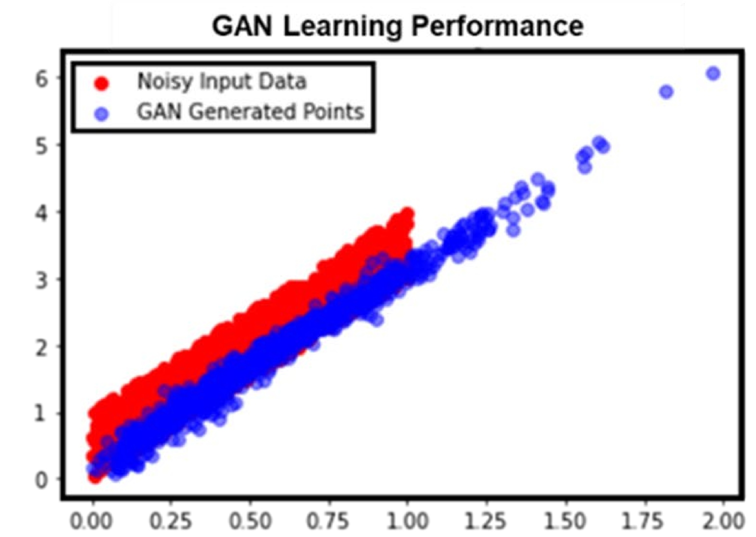


Figure 4: GAN Learning Demonstration

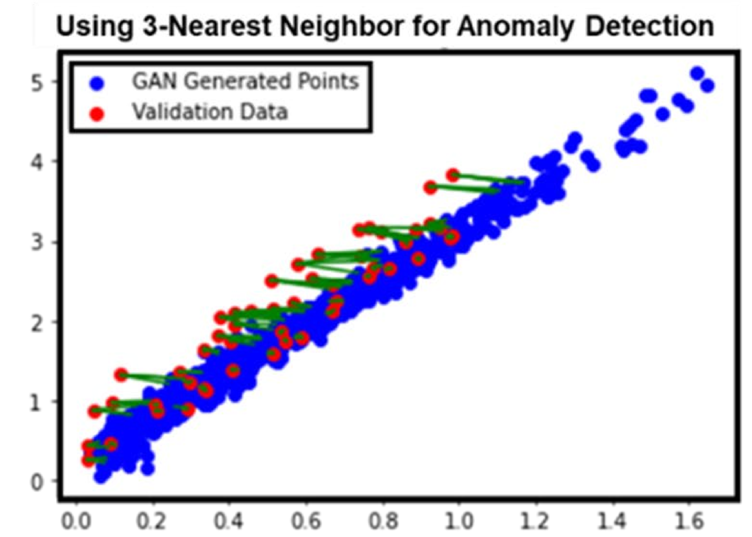


Figure 5: k-NN Demonstration

### Control Chart Monitoring Performance (Case 2)

| Control limit | False Positive Rate (%) | Anomaly Detection Rate (%) | Out of Control ARL |
|---|---|---|---|
| 95th Quantile | 26.41 (1.02) | 96.82 (0.58) | 1.1 (0.32) |
| 99th Quantile | 1.52 (0.40) | 87.65 (1.14) | 1.1 (0.32) |
| $\bar{\bar{d}} + 2\sigma$ | 16.92 (0.85) | 95.53 (0.74) | 1.1 (0.32) |
| $\bar{\bar{d}} + 3\sigma$ | 4.55 (0.75) | 91.48 (0.95) | 1.1 (0.32) |

### Control Chart Monitoring Performance (Case 3)

| Control limit | False Positive Rate (%) | Anomaly Detection Rate (%) | Out of Control ARL |
|---|---|---|---|
| 95th Quantile | 69.39 (1.62) | 99.61 (0.15) | 1.0 (0.00) |
| 99th Quantile | 24.99 (1.14) | 95.54 (0.63) | 1.1 (0.32) |
| 99.5th Quantile | 15.89 (0.96) | 92.42 (0.68) | 1.2 (0.42) |
| $\bar{\bar{d}} + 2\sigma$ | 65.78 (1.71) | 99.48 (0.26) | 1.0 (0.00) |
| $\bar{\bar{d}} + 3\sigma$ | 36.69 (1.21) | 97.53 (0.59) | 1.1 (0.32) |
| $\bar{\bar{d}} + 4\sigma$ | 15.55 (0.87) | 92.27 (0.71) | 1.2 (0.42) |
| $\bar{\bar{d}} + 5\sigma$ | 4.91 (0.61) | 82.54 (0.80) | 1.2 (0.42) |
| $\bar{\bar{d}} + 6\sigma$ | 1.51 (0.35) | 68.15 (1.43) | 1.5 (0.85) |

## Conclusion and Future Work

- **GAN** successfully **captures distribution** and **eliminates noise**
- Simulation study demonstrates **promising performance** to deal with data streams with various dimensionalities and magnitudes
- Current research focusses on early-stage proof-of-concept
- More comprehensive experiments with more complicated anomalies are left for future research

## References


- **Slater, K.**, Li, Y., Wang, Y., Shan, Y., and Liu, C., 2023, "A Generative Adversarial Network (GAN)-Assisted Data Quality Monitoring Approach for Out-of-Distribution Detection of High Dimensional Data," Proceedings of 2023 IISE Annual Conference. (Accepted)
- Goodfellow, I., *et al*., 2020, "Generative adversarial networks," Communications of the ACM, 63(11), 139-144.


## Acknowledgement


- This work is supported by the National Science Foundation (NSF) under Award Number IIP-2141184.



**PI Contact Information: Dr. Chenang Liu, email: Chenang.Liu@okstate.edu, phone: 405-744-6055**