

대한산업공학회
제 16회 석사논문 경진대회

1차원 합성곱 신경망과 적대적 생성 신경망을 활용한 실시간 이상 감지

RFDGAN: Real-Time Fault Detection Model Using 1D CNN and GAN

송승환¹, 백준걸^{1*}

¹고려대학교 산업경영공학과

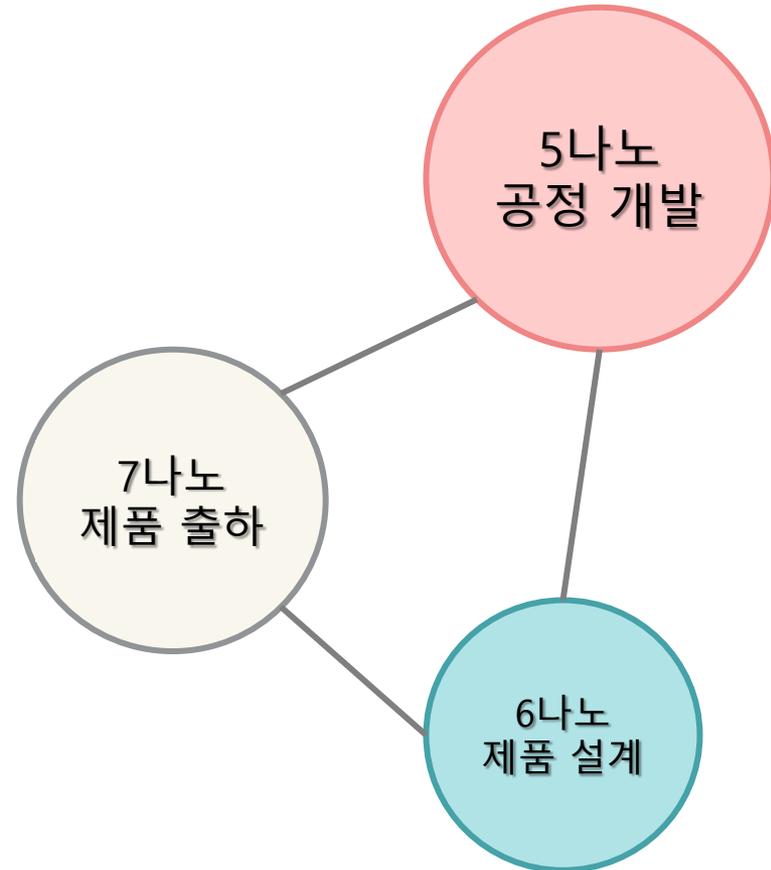
{ss-hwan, jungeol}@korea.ac.rk

목차

- 01 연구 배경
- 02 관련 연구
- 03 제안 방법론
- 04 실험 결과
- 05 결론

[01] 연구 배경

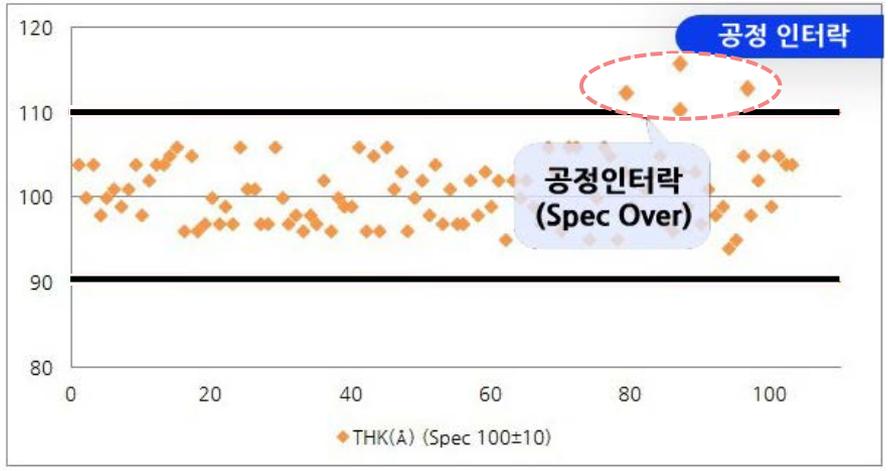
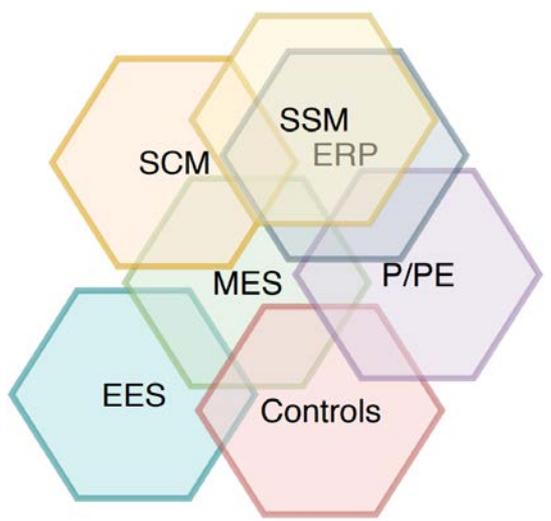
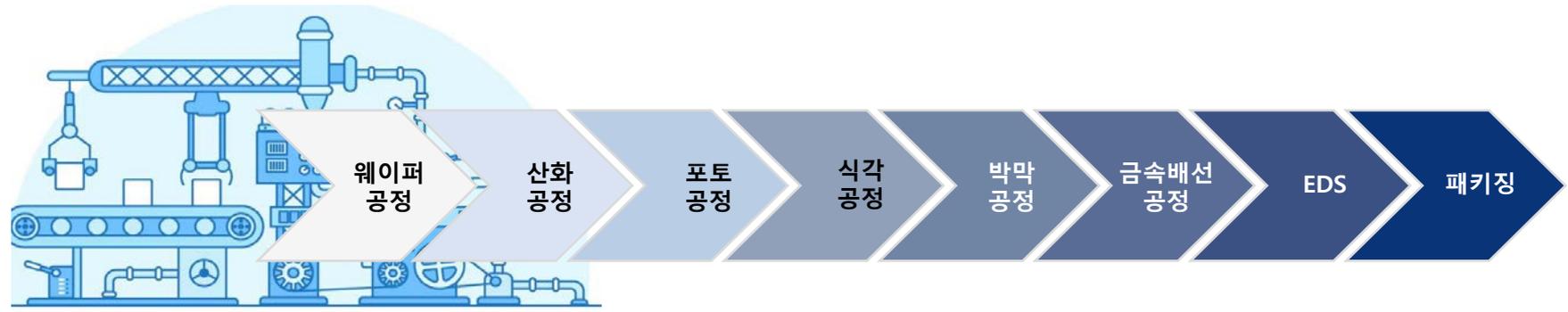
【 4차 산업혁명과 반도체 제조 공정 】



INDUSTRY 4.0: 초미세화 공정 개발 → 정밀하고 자동화된 공정 운영

[01] 연구 배경

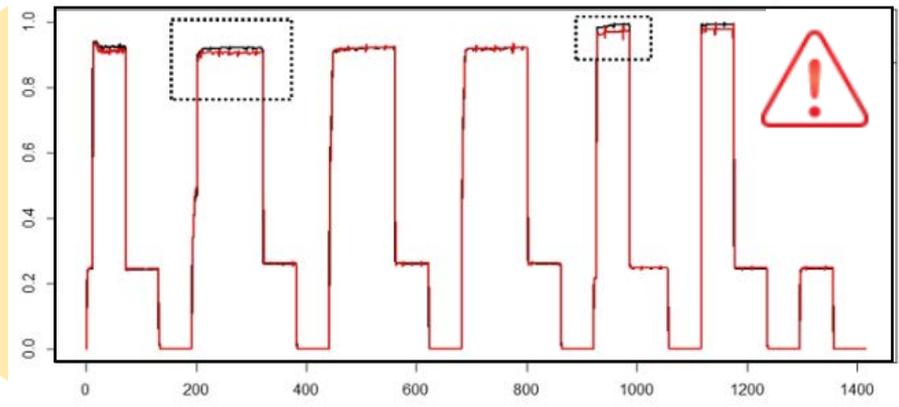
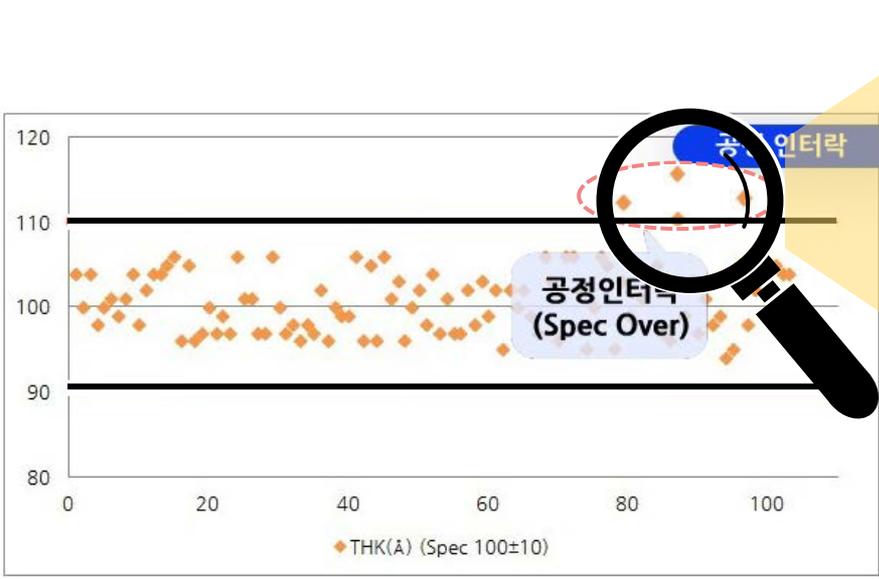
【 실시간 이상 감지의 필요성 】



- 정상 범위를 나타내는 인터락 조건을 넘어가면 알람 발생으로 사전 조치 수행

[01] 연구 배경

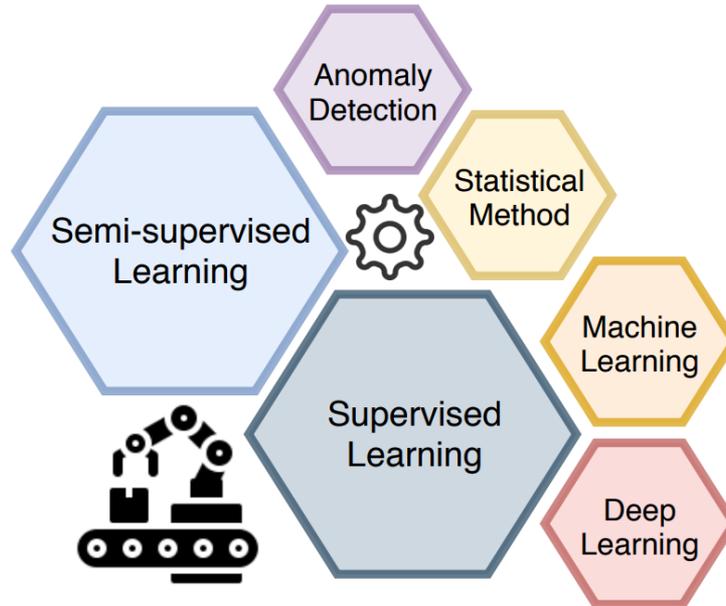
【 인터락 유발 데이터의 진/가성 분류 】



진/가성 불량 분류 ~~***~~

- 최근에는 보다 정밀한 FDC 시스템을 구축하기 위해 단순히 정상과 비정상을 구분하는 것이 아니라, 비정상 데이터로 판별된 데이터에 대한 심층적 분석이 요구됨
- 인터락 유발 데이터의 진/가성 분류 문제는 두가지 현실적인 문제를 가짐
 - 1) 인터락을 유발시킨 진/가성 데이터는 서로 극미한 차이만 보이는 경우가 대다수
 - 2) 진성 데이터 발생 빈도가 현저히 낮아, 사용할 수 있는 데이터가 거의 존재하지 않음

【 반도체 공정에서의 FDC 시스템 】



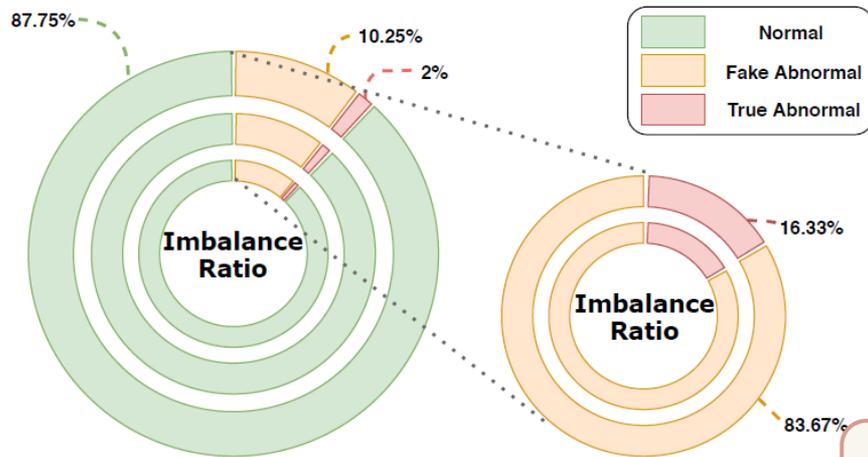
- **통계적(Statistical) 방법**
 - 주로 공정 매개변수를 이용한 관리도 방법으로 사용됨
- **기계 학습(Machine learning) 방법**
 - 주로 특징 추출에 인공지능 방법을 결합한 방법으로 사용됨
 - Real-Time Contrast을 통해 실시간 공정 모니터링에 적용됨
- **심층 학습(Deep learning) 방법**
 - 통계적 방법과 기계 학습 방법보다 성능이 뛰어남
 - 특히, 복잡한 특징 추출 과정을 모델 내에서 진행 가능

한계점

- 모델 설정과 공정 매개변수 선정의 어려움이 존재
- 전체 신호를 대표하는 특징의 품질이 낮은 경우 성능이 제한적
- 계산 복잡성이 높고, 충분한 양의 데이터가 필요

[02] 관련 연구

【 반도체 공정에서의 불균형 문제 】



한계점

• Random Oversampling

- 소수 클래스의 인스턴스를 임의로 복제하여 수를 늘리는 방법 (단순 복사)

• Synthetic Minority Oversampling Technique

- 소수 클래스의 데이터를 단순하게 복제함으로써 생기는 과적합을 피하기 위해 사용
- 소수 클래스 샘플들을 중심으로 k-최근접 이웃을 이용하여 새로운 합성 샘플을 생성

• Generative Adversarial Network

- 데이터를 분포에 기반하여 생성하는 방법
- 주로 이미지 생성 분야에서 우수한 성능을 보였으나, 최근 신호 데이터에 대해서 우수한 성능을 보임

→ 정보의 손실이 없지만 소수 클래스를 단순 복사하기 때문에 **과적합의 가능성이 있음**

→ 데이터를 생성하는 동안 인접해 있는 다수 클래스 데이터의 위치를 고려하지 않기 때문에 **클래스가 겹치거나 노이즈를 만들 수 있으며, 실제 데이터의 분포를 이용하지 않는 문제가 존재**

→ 기울기 소실(**vanishing gradient**), 모드 붕괴(**mode collapse**) 문제 존재

[03] 제안 방법론

【 제안 방법론 이점 】



정확성

- ✓ 입력 데이터가 불완전하거나 노이즈가 있어도 잘 동작하는 CNN을 사용하여, 복잡한 특징 추출 과정을 모델 내에서 진행
- ✓ 오버샘플링 방법들의 단점을 보완하여, 시계열 데이터 생성에 우수한 성능을 보임



신속성

- ✓ 최적의 구조를 실험적으로 찾아냄으로써, CNN의 많은 파라미터 수로 인한 높은 계산 복잡성 문제를 해결
- ✓ 데이터 불균형 문제 해소와 동시에 분류를 수행하는 end-to-end 방법으로, 빠른 처리가 요구되는 실시간 탐지에 적합

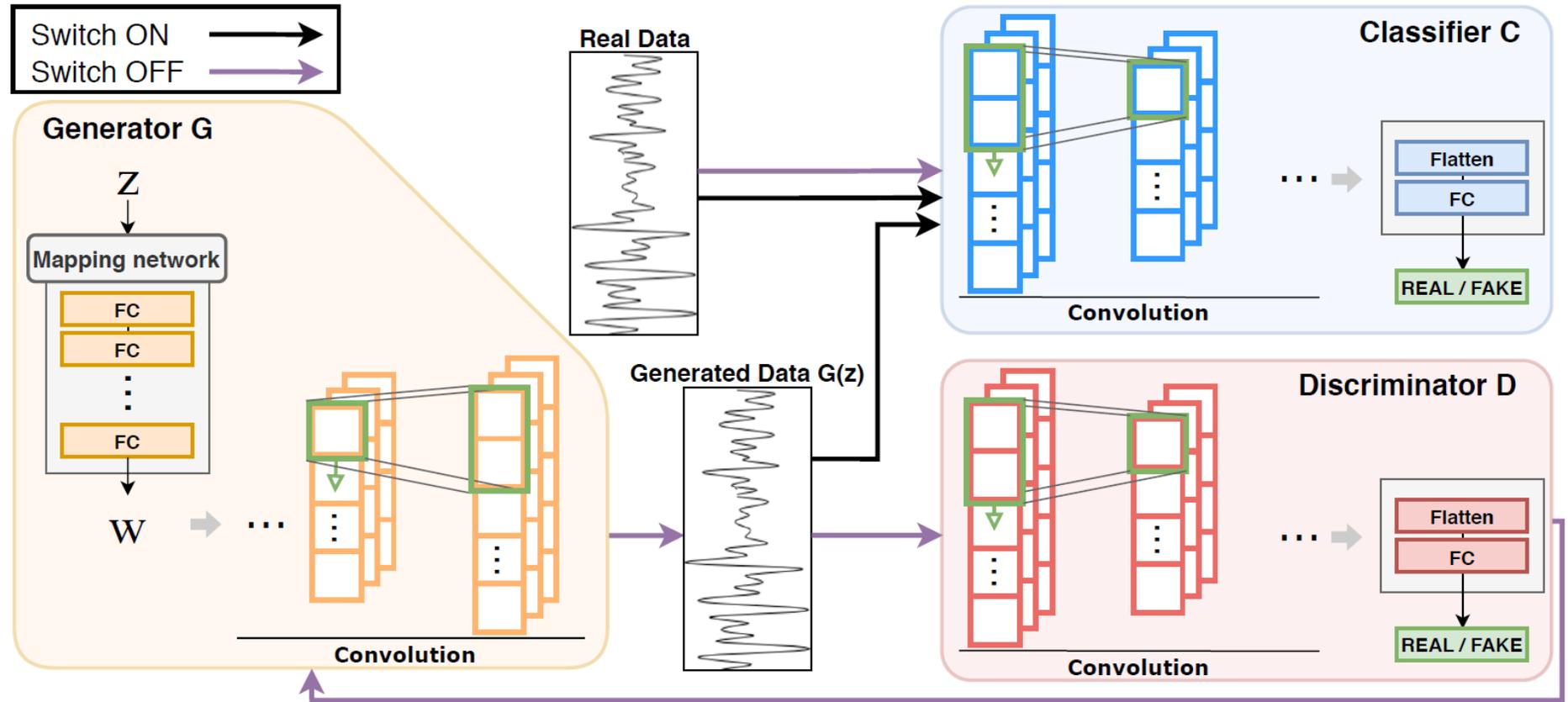


경제성

- ✓ 수집 비용이 낮은 원본 데이터를 사용하며, 시계열 데이터를 이미지 변환 없이 사용하므로 원본 데이터 정보 손실이 없음
- ✓ 인터락 유발 데이터의 진/가성 분류를 통해 불필요한 알람을 줄이고 후속 정비 비용을 절감하여, 생산성과 수율이 향상됨

[03] 제안 방법론

【 제안 방법론 구조 】

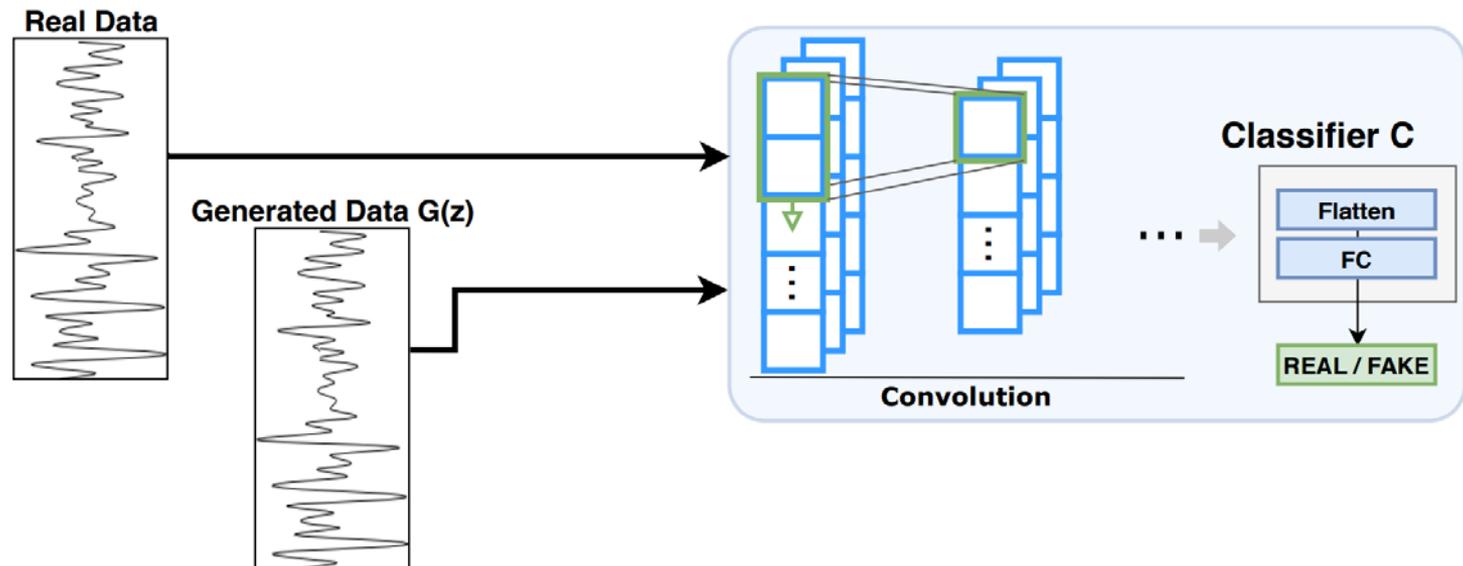


[03] 제안 방법론

【 제안 방법론 구조 】

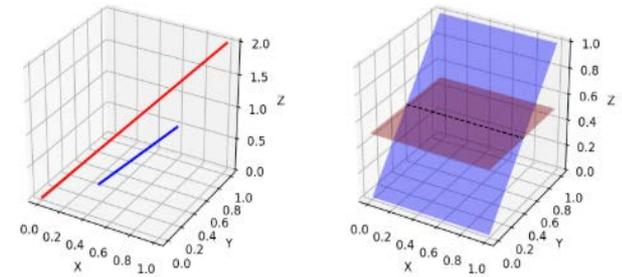
❖ 1D CNN

- 본 연구에서 사용된 CNN은 1D CNN으로, 필터가 한 방향으로만 움직이면서 합성곱(convolution) 연산이 수행됨
- 1D CNN은 복잡한 구조를 사용하지 않아도 우수한 성능을 보이며, 주로 시계열 데이터의 실시간 변화를 알기 위해 사용됨
- 1D CNN 구조를 통해 시간에 따른 인터락 유발 진/가성 데이터 변화 반영 가능



[03] 제안 방법론

【 제안 방법론 구조 】



고차원 공간에서 저차원 매니폴드

❖ 1D GAN – Loss Function

- 본 연구에서는 GAN의 기울기 소실(vanishing gradient)과 모드 붕괴(model collapse) 문제를 해결하기 위해 거리지표로 1-Wasserstein을 사용
- Wasserstein vs JS or KL divergence
 - 두 분포가 유니폼 분포를 따르고, θ 가 0이 아니면 P 와 Q 는 겹치지 않는다고 할 때,
 - θ 가 0이 아닐 때, D_{KL} 는 무한대 값을 갖게 되고, D_{JS} 값이 갑자기 튀게 되어 미분이 불가능해짐
 - 저차원 매니폴드에서 두 분포가 겹치지 않을 때, Wasserstein distance는 여전히 의미 있는 값을 가지며 연속적으로 표현됨

KL (Kullback-Leibler) Divergence

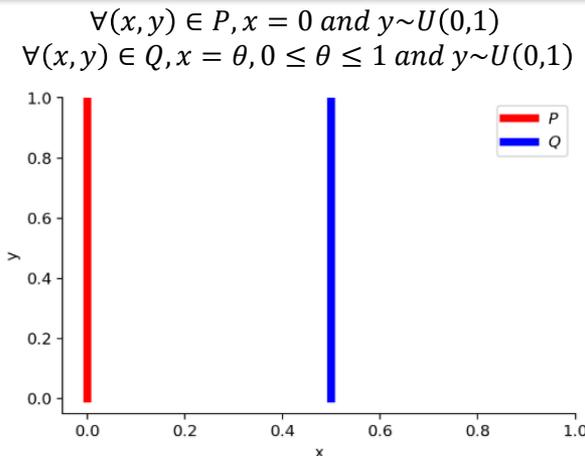
- p 분포가 다른 분포 q 와 얼마나 떨어져 있는지 측정
- $D_{KL}(p||q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx$

JS (Jensen-Shannon) Divergence

- 두 분포의 유사도를 측정하는 metric으로 [0, 1] 사이의 값을 갖으며 대칭적인 구조
- $D_{JS}(p||q) = \frac{1}{2}D_{KL}(p||\frac{p+q}{2}) + \frac{1}{2}D_{KL}(q||\frac{p+q}{2})$

Wasserstein Distance

- 어떤 확률 분포 모양을 띄는 흙더미를 다른 확률 분포 모양을 가지도록 하는데 드는 최소 비용
- $W(p_r, p_g) = \inf_{\gamma \in \Pi(p_r, p_g)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|]$, $\Pi(p_r, p_g)$



when $\theta \neq 0$:

$$D_{KL}(P||Q) = \sum_{x=0, y \sim U(0,1)} 1 * \log \frac{1}{0} = +\infty$$

$$D_{KL}(Q||P) = \sum_{x=\theta, y \sim U(0,1)} 1 * \log \frac{1}{0} = +\infty$$

$$D_{JS}(P||Q) = \frac{1}{2} (\sum_{x=0, y \sim U(0,1)} 1 * \log \frac{1}{1/2} + \sum_{x=\theta, y \sim U(0,1)} 1 * \log \frac{1}{1/2}) = \log 2$$

$$W(P, Q) = |\theta|$$

when $\theta = 0$:

$$D_{KL}(P||Q) = D_{KL}(Q||P) = D_{JS}(P||Q) = 0$$

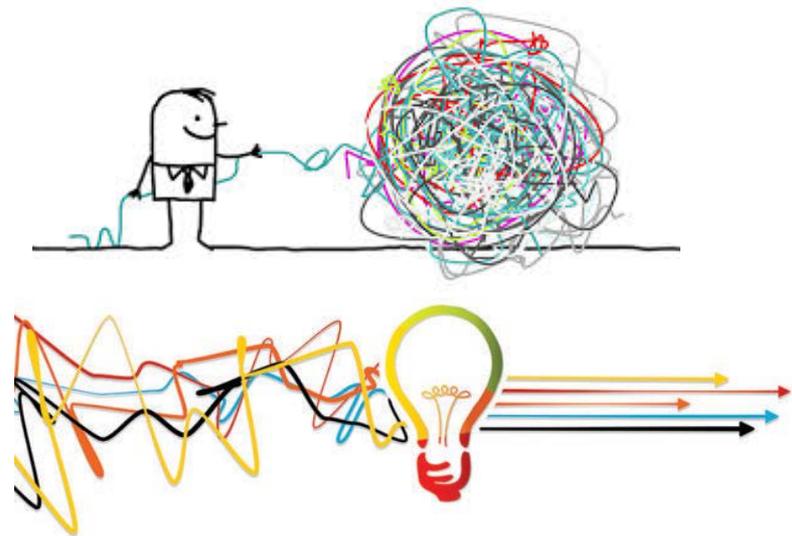
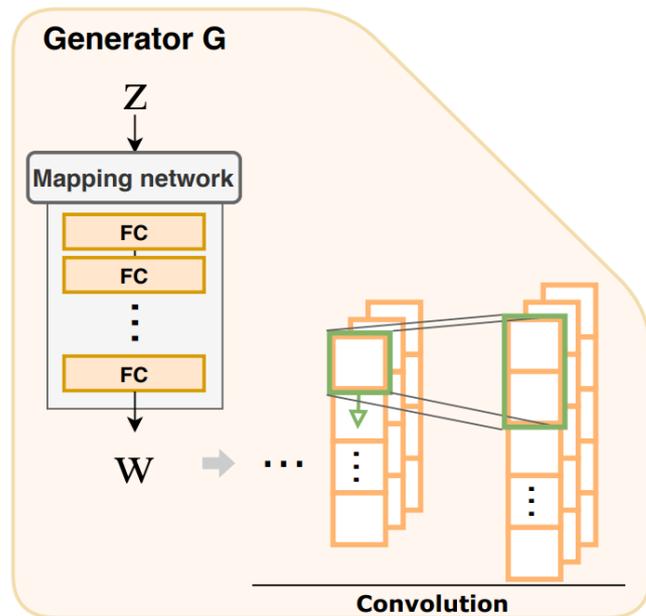
$$W(P, Q) = 0 = |\theta|$$

[03] 제안 방법론

【 제안 방법론 구조 】

❖ 1D GAN – Style based Generator

- 본 연구에서는 데이터 생성의 품질을 높이고자 StyleGAN의 구조를 1D GAN generator로 사용
- 잘 학습된 GAN 모델은 잠재 변수가 선형 부분공간을 가지게 되고, 변화에 대한 직접적인 조절이 가능해짐
- 잠재변수 z 가 mapping network f 를 거친 w 가 generator의 입력값으로 들어가기 때문에 학습 데이터의 분포를 정확히 맞추며, 생성된 데이터의 품질이 올라가게 됨



[03] 제안 방법론

【 제안 방법론 학습 과정 】

❖ Phase 1 – Switch OFF

- Generated data의 품질이 좋지 않기 때문에 switch를 off !
- 시계열 데이터에 대해 1D CNN Classifier를 이용한 pre-training 수행 (overfitting 방지)

$$\text{Binary Cross Entorpy} = -\frac{1}{N_p} \sum_{i=0}^{N_p} y_i \cdot \log(f_c(y_i^r)) + (1 - y_i) \cdot \log(1 - f_c(y_i^r))$$

- Discriminator는 실제 데이터와 생성된 데이터를 구분하는 방법을 학습
- Generator는 discriminator가 구분하지 못하도록 생성 품질을 높이는 것을 목적으로 함
- Discriminator와 generator의 경쟁 학습을 통해 GAN 모델이 학습됨
- WGAN-GP loss function을 적용하여, 안정적인 학습을 수행

$$L = \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g} [D(\tilde{x})] - \mathbb{E}_{x \sim \mathbb{P}_r} [D(x)] + \lambda \mathbb{E}_{\tilde{x} \sim \mathbb{P}_{\tilde{x}}} [(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2]$$

❖ Phase 2 – Switch ON

- 실제 데이터와 생성된 데이터 모두를 이용하여 1D CNN Classifier를 업데이트

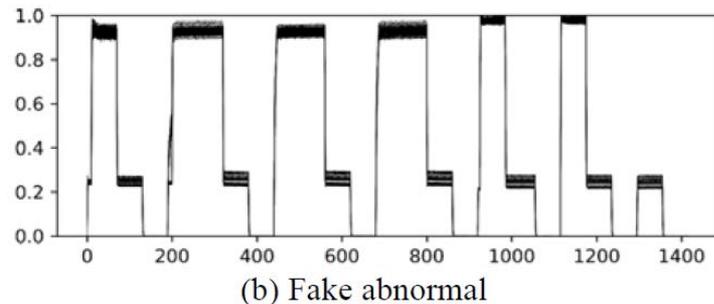
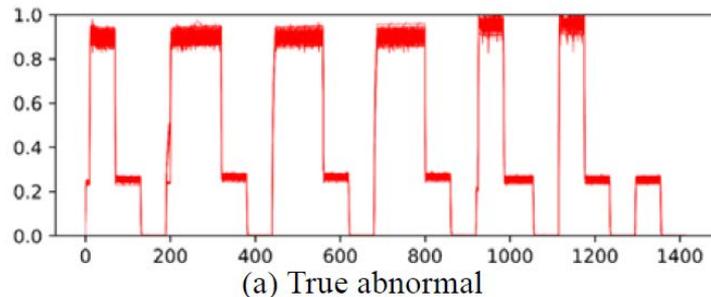
$$\text{Binary Cross Entorpy} = -\frac{1}{N} \sum_{i=0}^N y_i \cdot \log(f_c(y_i^a)) + (1 - y_i) \cdot \log(1 - f_c(y_i^a))$$

[04] 실험 결과

【 데이터 수집 및 실험 설계 】

❖ 데이터 수집

- 본 연구에서 사용된 데이터는, 웨이퍼 표면에 고농도의 불순물을 고온(900~1,000°C) 으로 투입시키는 확산 공정 과정에서 수집된 실제 데이터를 사용
- 두 데이터가 서로 유사한 모습을 띄기 때문에, 정확한 특질 추출은 필수적임



❖ 실험 설계

- Data length 가 다른 file 삭제 (1295로 통합)
- Input_height= 1295, filter_height= 100, stride= 10 (진성 100개, 가성 600개), 10-fold
- 데이터 셋을 선행(pilot):학습(train):평가(test)=1:6:3 비율로 분할

1단계 실험) 학습 데이터만을 통하여 1D CNN과 1D GAN에 대한 최적의 구조를 찾는 실험을 진행

2단계 실험) 전체 데이터셋을 통하여 RFDGAN의 성능 평가를 위한 실험을 진행

[04] 실험 결과

【 성능 평가 지표 】

❖ 분류 성능 평가 지표

Precision

$$\frac{TP}{TP + FP}$$

Recall

$$\frac{TP}{TP + FN}$$

Accuracy

$$\frac{TP + TN}{TP + FP + FN + TN}$$

F1 score

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

	Predicted Yes	Predicted No
Actual Yes	TP (True Positive)	FN (False Negative)
Actual No	FP (False Positive)	TN (True Negative)

❖ 생성 모델 평가 지표

Fréchet Inception Distance

$$d^2((m, C), (m_w, C_w)) = \|m - m_w\|_2^2 + \text{Tr}(C + C_w - 2(CC_w)^{1/2})$$

m : real data의 평균 C : real data의 공분산
 m_w : fake data의 평균 C_w : fake data의 공분산

[04] 실험 결과

【 최적 1D CNN 구조 설계 】

Optimal structure of 1D CNN

Number of layers	Filter size: 100		Filter size: 200	
	Acc	Time (s)	Acc	Time (s)
1	0.51	30.29	0.50	29.23
2	0.54	33.78	0.53	33.15
3	0.54	36.12	0.54	35.86
4	0.55	42.43	0.55	41.54

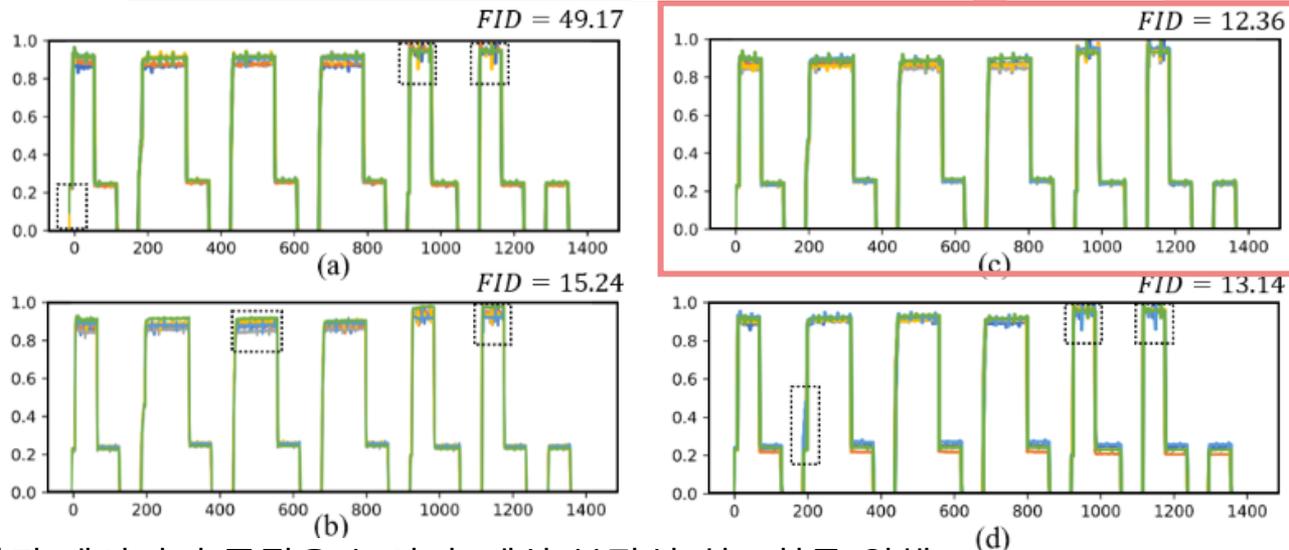
- CNN의 많은 파라미터 수로 인해 계산 복잡성이 높고, 데이터가 많이 필요하다는 단점을 해결하기 위해, 1D CNN의 구조를 활용
- 오버샘플링을 수행하지 않은 원본 데이터를 이용하여 실험을 진행하였으며, 데이터의 모양을 참고하여 필터 크기를 100과 200으로 실험을 진행
- 분류 정확도와 훈련에 소요된 시간을 고려하여 필터 크기는 100, layer는 2개 사용

[04] 실험 결과

【 최적 1D GAN 구조 설계 】

Generated result by mapping network(3 layers), (a) traditional (b) style-based 2, (c) style-based 4, (d) style-based 6

Method	Number of layers			
	1	2	3	4
Traditional	67.00	52.53	49.17	49.03
Style-based 2	45.21	31.21	15.24	14.98
Style-based 4	41.87	29.45	12.36	12.56
Style-based 8	40.55	26.20	13.14	12.49



- 생성된 데이터의 품질을 높이며, 계산 복잡성 최소화를 위해 mapping network f 와 1D GAN의 최적 성능 구조를 찾는 실험을 진행
- Layer를 3개, mapping network를 4개로 구성했을 때, 전체 모양은 유지한 채 분산에서만 차이가 존재

[04] 실험 결과

【 최적 1D GAN 구조 설계 】

Classification result for various generator designs

Method	Accuracy	Precision	Recall	F1 Score
A Baseline	0.54	0.00	0.00	0.00
B A + RO	0.69	0.50	0.73	0.59
C A + SMOTE	0.71	0.64	0.76	0.69
D A + GAN	0.86	0.89	0.83	0.85
E + DCGAN	0.91	0.89	0.86	0.87
F + WGAN-GP loss function	0.94	0.94	0.95	0.95
G + Add mapping network (RFDGAN)	0.99	0.99	0.98	0.99

- RO, SMOTE와 같은 기존 방법은 시계열 데이터를 정확히 생성해내지 못하는 단점이 있기 때문에 분류 성능이 떨어짐
- 실험을 통해 손실 함수로 WGAN-GP loss function을 사용하거나 mapping network를 추가한 결과가 전체적인 분류 성능에 좋은 결과를 보임
- 제조 공정은 주로 불량으로 예측한 데이터 중 실제 불량 비율을 나타내는 *precision*을 통해 모델에 대한 신뢰성을 확보하므로, RFDGAN이 우수한 성능을 보임을 알 수 있음

[05] 결론

【 연구 의의 및 개선점 】

- 본 연구는 인터락을 유발시킨 진/가성 데이터의 실시간 분류 절차를 고도화하기 위해 RFDGAN을 제안
- RFDGAN은 최적 1D GAN 구조를 이용하여 불균형 문제를 해결하며, 동시에 최적 1D CNN 구조를 이용하여 노이즈가 있는 상황에서도 자동으로 특질을 추출하여 분류를 수행
- RFDGAN을 통해 인터락을 유발시킨 진/가성 데이터를 탐지함으로써, 불필요한 알람을 줄이고 후속 정비(maintenance) 비용을 절감하여, 반도체 공정의 생산성과 수율 향상 기대

[다양한 분야 적용을 통한 강건성 확보]

- 불균형이 존재하는 다양한 공정에 적용 가능하므로 실용화 가능성이 크고, FDC가 아닌 다른 분야에도 적용 가능할 것으로 기대
- 향후 연구로 확산 공정 내 여러 센서에 대해 실험을 적용해보는 것을 통해, 여러가지 불균형 비율이 존재하는 상황에서 RFDGAN을 적용하는 다양한 시도와 적용이 모색됨

[모델 평가지표 개선과 해석 능력 향상]

- 주로 이미지 데이터에서 사용되는 FID를 시계열 데이터에 적용하는 것은 한계가 존재
- 진성 불량률 전부를 감지하여 불필요한 알람을 줄이고 생산성과 수율 향상에 기여하기 위해서는 precision의 성능을 더욱 개선시켜야 할 필요성이 있음

대한산업공학회
제 16회 석사논문 경진대회

Q & A

감사합니다

이 연구는 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원으로 수행된 연구(NRF-2019R1A2C2005949)이며, 2019년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임(P0008691, 2019년 산업전문인력역량강화사업).

Appendix

【 학습 과정의 수도 코드 】

Algorithm 1. RFDGAN 학습 과정

Input: The set of pilot data $y_i^r \in D_{pilot}$,

the set of train data D_{train} ,

the size for the real abnormal $num(real)$,

the size for the fake abnormal $num(fake)$,

the switch variable S

the generated data quality until i^{th} batch, FID_i

the generating quality threshold h

the set of generated data $y_i^a \in D_{train}'$

Output: The detection accuracy for each true/fake abnormal data of the diffusion process

initialize the switch, $S = 0$

While $num(real) < num(fake)$ and $FID_i > h$ **do**

pre-train the 1D CNN model using y_i^r in D_{pilot}

train the 1D GAN model using D_{train}

if $num(real) = num(fake)$ **then**

$S = 1$

update the 1D CNN model using y_i^a

end

$i = i + 1$

End

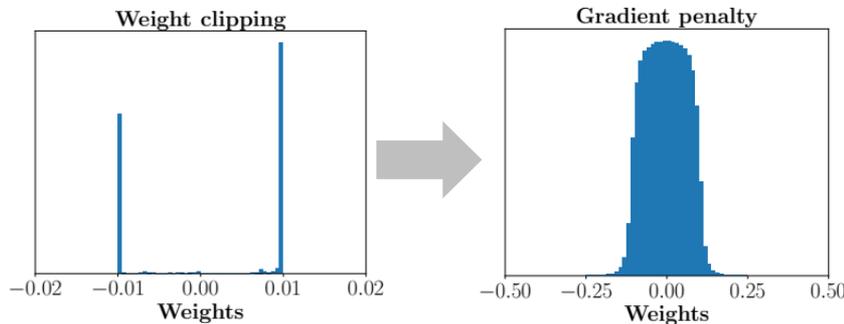
- 진/가성 데이터 수가 같아지거나, 특정한 종료 조건을 만족할 때까지 반복적으로 수행
- 특정 조건
 - FID가 데이터 생성에 대한 특정 조건인 h 보다 작거나 같을 때
 - i 번째 FID와 $i - 1$ 번째 FID의 차이가 거의 없을 때

Appendix

【 WGAN-GP weight clipping 】

- Diminished capacity. Lots of Lipschitz functions are not included in this family
- Exploding or vanishing gradients depending on c
- Weights tend to saturate i.e. $-c$ or c
- Unstable with momentum-based SGD techniques

In general, there are problems with weight clipping for Lipschitz condition.



$$\arg \max_{\|f\|_L \leq 1} \mathbb{E}_{y \sim \mathbb{P}_r} [f(y)] + \mathbb{E}_{x \sim \mathbb{P}_g} [f(x)]$$

$$\text{Optimal } f^*, \quad \|\nabla f^*(x_t)\| = 1$$

$$L = \underbrace{\mathbb{E}_{\tilde{x} \sim \mathbb{P}_g} [D(\tilde{x})] - \mathbb{E}_{x \sim \mathbb{P}_r} [D(x)]}_{\text{Original critic loss}} + \lambda \underbrace{\mathbb{E}_{\tilde{x} \sim \mathbb{P}_{\tilde{x}}} [(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2]}_{\text{Our gradient penalty}}$$

Original critic loss

Our gradient penalty