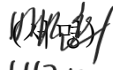
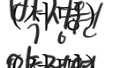
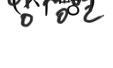


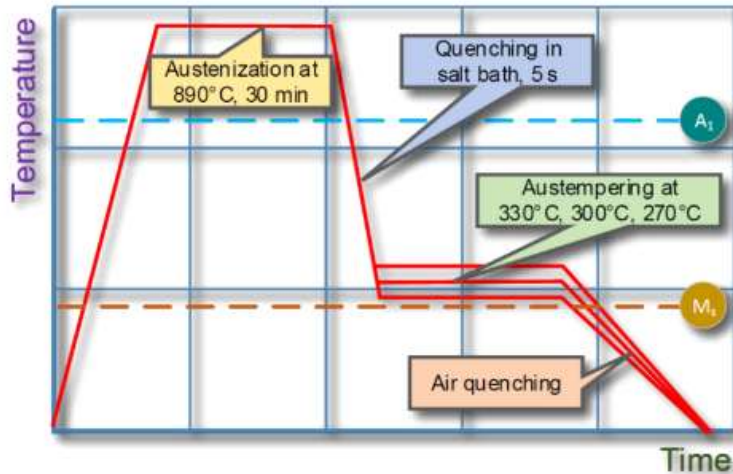
# 제3회 K-인공지능 제조데이터 분석 경진대회 보고서

프로젝트명	딥러닝 기반 열처리 공정 불량 예측 및 요인 분석
팀명	Absolute A
내용요약	<p>제조 공정의 효율성을 높이고 불량률을 감소시키는 것은 기업에게 있어 중요한 과제이다. 특히 기술적으로 복잡하며 고가의 원료를 사용하는 열처리 공정의 경우, 불량 제품의 발생은 손실을 크게 증가시키므로 이를 예방하는 것이 중요하다. 그러나 실제 제조 현장에서는 공정 중에 제품의 품질을 실시간으로 평가하는 것이 어렵고, 비용과 기술적 제약으로 모든 제품을 검사하는 것이 현실적이지 않다. 이 문제를 해결하기 위해 제안된 분석 방법론은 공정 후 전체 불량률에 기반해 라벨링한 뒤 해당 공정의 시간 표현을 학습하는 접근을 제시한다. 각 공정의 공정 시간과 분당 생산량이 일정하지 않고 같은 시간대의 로그가 다른 제품을 나타낼 수 있다는 현실을 반영하여, 공정이 완료된 후의 전체 불량률을 바탕으로 라벨링을 진행한다. 임계값 이상의 불량률을 기준으로 '불량' 또는 '정상'으로 라벨링 한 뒤 해당 공정의 시계열을 학습한다.</p> <p>분류 작업은 LSTM, GRU, TCN, Transformer 등과 같은 시계열 딥러닝 모델을 이용하여 진행되며, 이 모델들은 시계열 데이터로부터 독립적으로 특성을 추출하고 종속 변수와 독립 변수 간의 복잡한 관계를 학습한다. 최적의 모델이 선정된 후에는 shapley value 분석을 통해 해당 모델이 불량으로 분류한 요인 분석을 진행한다.</p> <p>본 프로젝트의 방법론을 통해 딥러닝 모델에서 도출된 변수의 중요도(shapley value 기반)를 분석하여, 이를 공정 책임자의 전문 지식과 접목시킬 수 있다. 딥러닝 모델이 제공하는 데이터 기반의 통찰과 책임자의 경험 및 직관을 결합함으로써, 더 정교하고 균형 잡힌 의사결정이 가능하다. 또한, 실시간 데이터 분석을 가능하게 함으로써, 공정 중 발생하는 문제에 신속히 대응이 가능하다.</p> <p>종합적으로, 제조업의 품질 관리에 있어서 데이터 기반 의사결정의 중요성을 강조한다. 특히, 딥러닝 예측 모델을 활용함으로써, 잠재적인 문제에 더욱 신속하고 정확하게 대응할 수 있다. 이는 품질 유지 및 생산 효율 향상에 직접적으로 기여하며, 결과적으로 비용 절감 및 운영 효율성 증진으로 이어질 수 있다는 기대효과를 제시한다.</p>
<p>상기 본인(팀)은 위의 내용과 같이 제3회 K-인공지능 제조데이터 분석 경진대회 결과 보고서를 제출합니다.</p> <p>2023 년 11월 3일</p> <p>팀장 : 배 소 희 </p> <p>팀원 : 박 정 원 </p> <p>팀원 : 양 정 열 </p> <p>한국과학기술원장 귀중</p>	

## □ 문제정의

### ○ 설비 개요

- austempering이란 금속 열처리 과정의 일종으로 주로 강철과 철 주물에 적용되어 특별한 기계적 성질을 부여한다.



열 처리 공정에 따른 온도 변화

- austempering 열처리 공정 단계<sup>[1]</sup>

- Austenization : 금속이 오스테나이트 상태가 될 정도로 고온으로 가열하는 단계이다. 금속은 일정 시간 동안 그 온도에서 유지되어 구조가 균일하게 될 때 유지한다.
- Quenching in salt bath : 염 욕조(salt bath)에 Austempering 온도까지 금속을 빠르게 냉각한다.
- Austempering : 오스테나이트를 바이나이트 구조로 변환한다.
- Air quenching : 금속은 공기 중에서 냉각하여 건조한다.

\*오스테나이트 : 철-탄소 합금에서 발견되는 비자성의 입방체적중심구조, 고온에서만 안정하며 상온에서는 대부분 불안정한 특성의 구조, 탄소를 상대적으로 높은 온도에서 용해할 수 있음

\*바이나이트 : 철-탄소 합금의 중간 미세구조

- Austempering 공정은 가열 유지시간이 짧아 탈탄 및 변형 발생이 매우 낮고, 금속의 인장 강도가 매우 우수하다는 장점이 있다. 또한, 균열(크랙)발생이 적어 표면 상태가 깨끗하여 후처리가 용이하다.<sup>[2][3]</sup>

- 그러나 대량 생산에 불리하며, 두께가 두꺼운 제품의 경우 경도값이 불균일한 단점이 있다. 따라서 두께가 얇고 크기가 작은 금속제품에 적합한 열처리 방법이다.
- 이 열처리 공정은 자동차 산업의 구동축 부품, 기계부품, 농업용 기계 등과 같이 충격 부하와 마모에 대한 저항력이 필요한 부품을 생산할 때 사용된다.

○ 이슈 사항(pain point)

- 제조 공정에서 설비 고장은 불량 제품의 증가, 설비 수리 기간 동안의 생산 중단, 설비 공정 수리 비용 등 기업 입장에서 천문학적 이익 감소로 이어진다.
- 열 처리 공정의 경우 제품 설계와 가공에 있어서 기술적 어려움이 많고, 사용되는 원료 자체도 공구강, 고속도강, 금형용강 등으로 고가이기 때문에 불량품이 발생할 경우 손실이 크다.
- 사전에 불량률을 예측하여 추가적인 불량품을 생산을 방지하는 것이 중요한 과제이다.
- 하지만 제품의 품질은 공정 진행 중 측정/평가가 어려워 공정 완료 후 진행하는 경우가 많아 공정 중 제품의 품질 수준을 실시간으로 평가하기 어렵다.
- 비용 문제로 공정 과정에서의 모든 제품의 품질을 측정하는 것 또한 불가능하여 샘플링을 하여 측정하기 때문에 정확한 품질 측정이 불가능하다.
- 따라서 공정 과정 중간에 품질 검수를 활용한 불량률 예측은 불가능하다.
- 또한, 장비 운용을 경험에 의존하는 경향이 있어, 이로 인해 운영의 효율성과 최적화가 저해되고 있다.
- 장비 운용을 데이터 기반의 근거에 의거하여 수행함으로써 경험에만 의존하는 작업 방식에서 벗어나 통합적인 의사결정 체계를 구축해야 한다.

○ 분석 목표

- 이슈 사항들을 보완하고 품질 유지 및 생산 효율 향상에 기여하는 분석 방법론을 제시한다.
- 설비 상태 데이터를 이용한 해당 공정에서의 불량 공정 시간 표현을 실시간으로 탐지하여 신속한 조치를 가능하게 할 수 있는 모델을 구축한다.
- 불량 공정 시간 표현이 탐지 되는 시계열에 대해서 어떤 설비 또는 센서에 문제가 발생했는지 파악하는 요인 분석을 진행한다.

## □ 제조데이터 정의 및 처리과정

○ 제조 데이터 정의

- 이용 공정 : austempering 열처리 공정
- 수집 기간 : 2022년 1월 ~ 2022년 7월
- 데이터 shape
  - ▶ data : (2939722, 21)
  - ▶ quality : (136, 7)
- 속성별 의미 및 type
  - ▶ data

변수명	의미	데이터 타입
TAG_MIN	데이터 수집 기간(단위 : 초)	object
배정번호	공정 작업 지시 번호	int64
건조 1~2존 OP	건조 온도 유지를 위한 출력량	float64
건조로 온도 1~2 zone	건조로 온도 zone의 온도	float64
세정기	세정기 온도	float64
소입1~4존 OP	각 소입존 온도 유지를 위한 출력	float64
소입로 CP 값	침탄 가스의 침탄 능력 량	float64
소입로 CP 모니터 값	소입로의 CP의 모니터링 값	float64
소입로 온도 1~4 zone	각 소입로 zone의 온도	float64
솔트 컨베이어온도 1~2 zone	솔트존 온도 유지를 위한 출력 값	float64
솔트조 온도 1~2 zone	각 솔트 zone의 온도값	float64

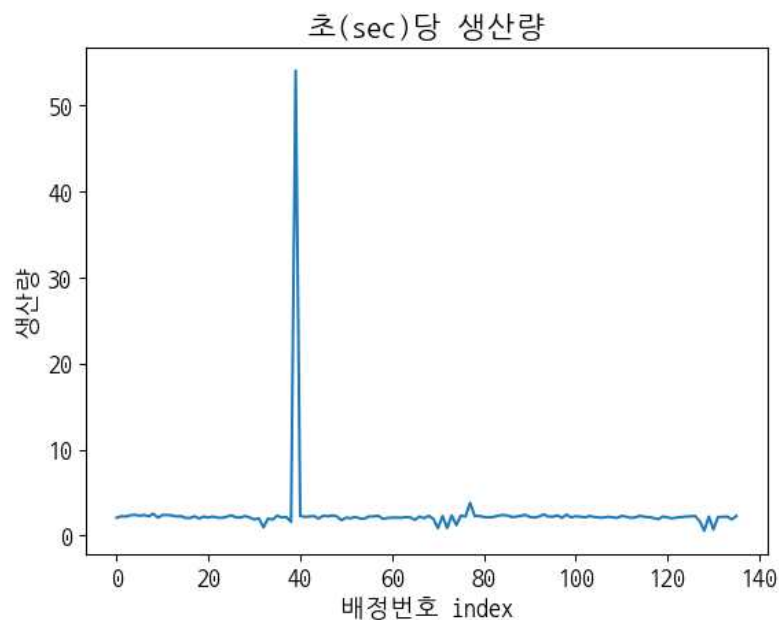
- \* 소입 : 담금질, 온도가 올라간 금속을 물이나 기름 등에 담가 급격하게 냉각시키는 작업
- \* 침탄법 : 탄소함유량이 0.2% 미만인 저탄소강이나 저탄소 합금강을 침탄제 속에 파묻고 오스테나이트 범위로 가열한 다음, 그 표면에 탄소를 침입하고 확산시켜 표면 층을 고탄소 조직으로 만드는 작업
- \* 솔트 : 염화물 또는 염화염, 열처리 공정에서 열을 전달하고 열처리 과정을 향상시키는 작업

▶ quality

변수명	의미	데이터 타입
배정번호	공정 작업 지시 번호	int64
작업일	공정 작업 날짜	datetime 64[ns]
공정명	공정 이름	object
설비명	설비 이름	object
양품수량	양품 생산 수량	int64
불량수량	불량 생산 수량	int64
총수량	전체 생산 수량	int64

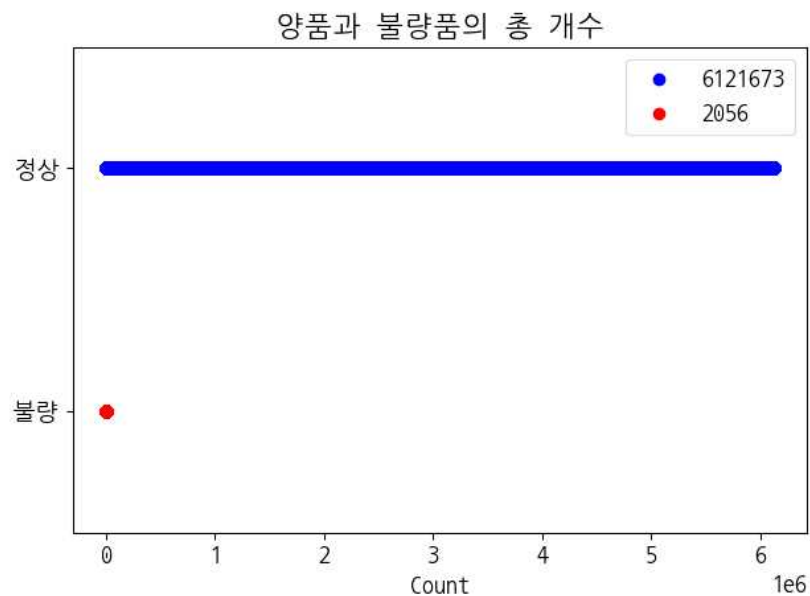
○ 데이터 주요 변수의 통계량 (EDA)

- 배정번호별 초당 생산량



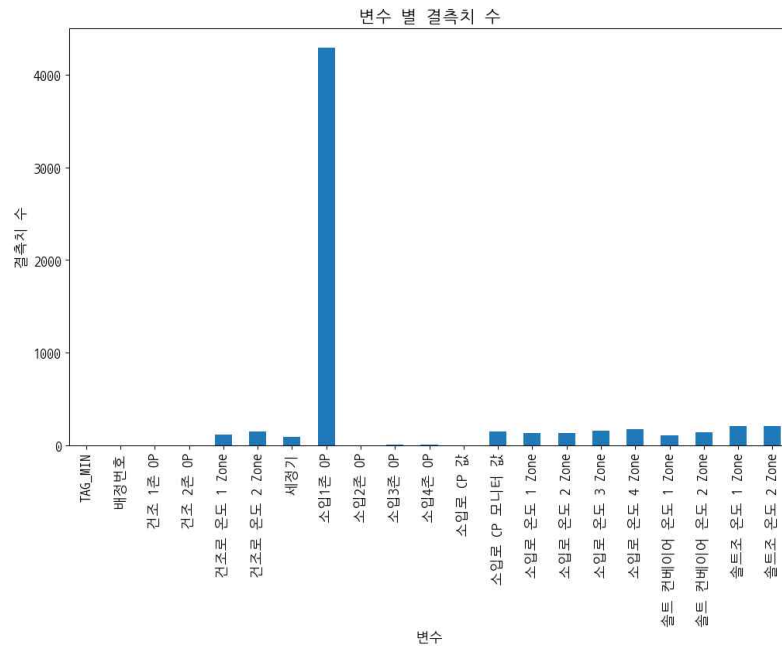
배정번호별 초당 생산량 그래프를 확인했을 때, 1초에 50개 이상을 생산하는 배정번호가 있는 것을 확인하였다. 확인해 본 결과 해당 배정번호는 실제로 초당 50개를 생산한 것이 아닌 data.csv에 공정 로그 데이터들이 기록되어있지 않는 것이다. 하지만 우리의 분석 방법인 불량공정과 정상공정의 부분 시간 표현을 학습하는 것에는 영향을 주지 않는 것으로 판단해 제거하지 않았다.

#### - 양품과 불량품 개수



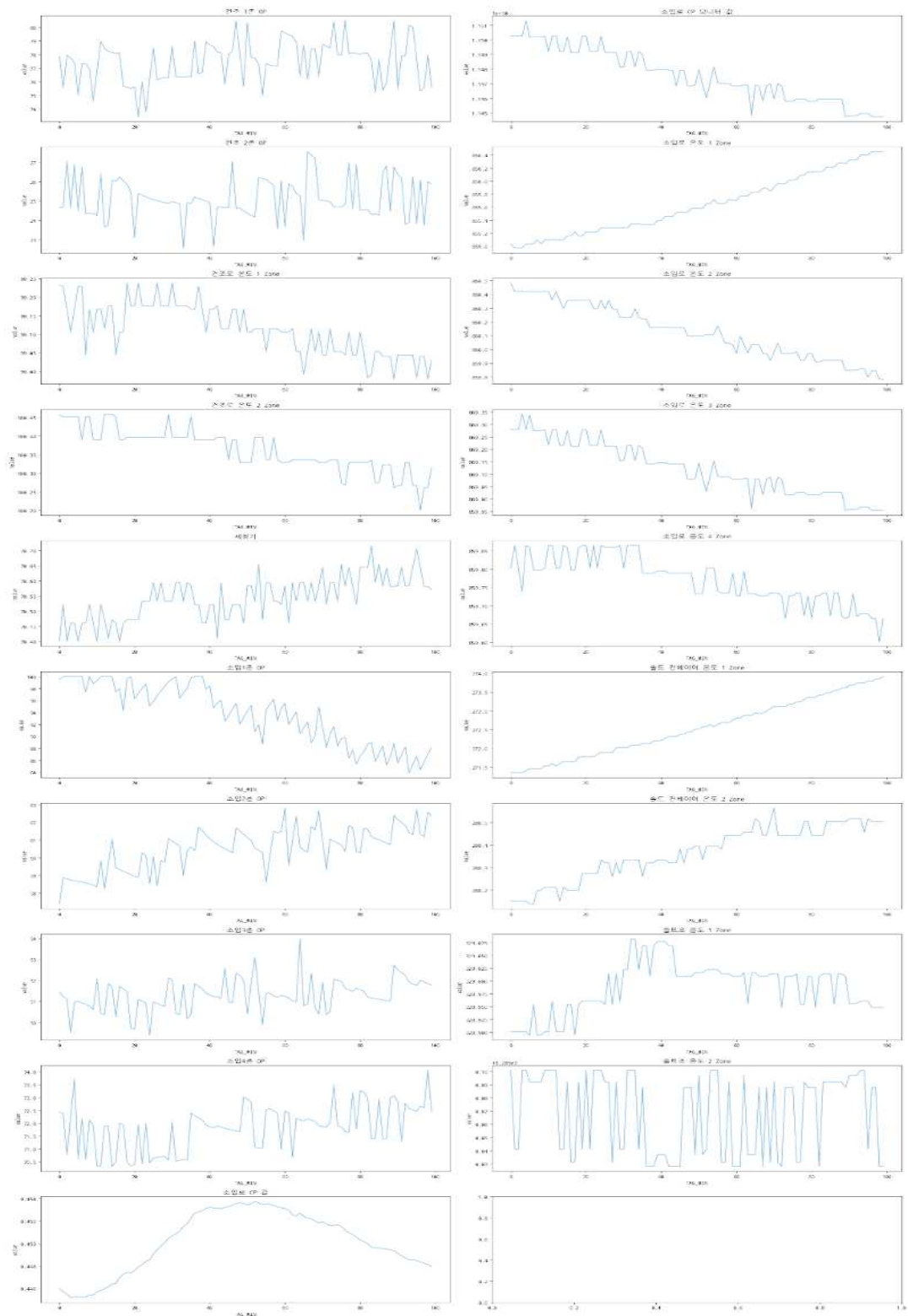
수집 기간 내 모든 생산품에 대해 양품의 개수는 6,121,673개인데 반해 불량품의 개수는 2,056개로 현저히 적은 수치이다. 양품 대비 불량품의 비율은 0.0003으로 양품 대비 불량품의 개수는 매우 적은 것을 알 수 있다.

## - 변수 별 결측치 개수



결측치가 가장 많은 소입 1존 op의 경우도 4000개 정도로 전체 2,939,722의 데이터에서 0.001의 비율로 결측치는 별다른 처리 없이 제거해도 분석 결과에 크게 영향을 미치지 않을 것으로 판단하였다.

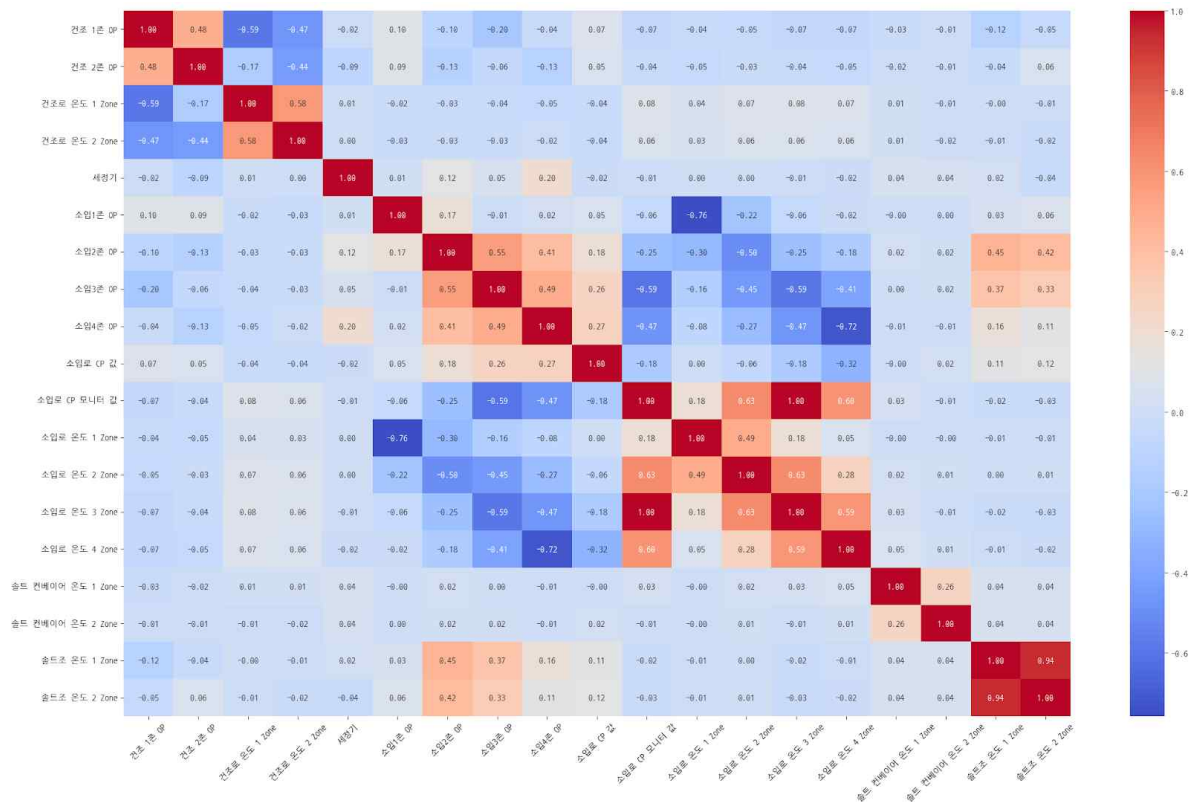
## - 각 변수에 대한 시계열 그래프



전체 변수에 대한 60초 간격의 시계열을 확인하였을 경우 비슷한 공정에서의 시계열은 유사하게 나타나나 특별한 계절성, 주기성을 확인하기 어려웠다.



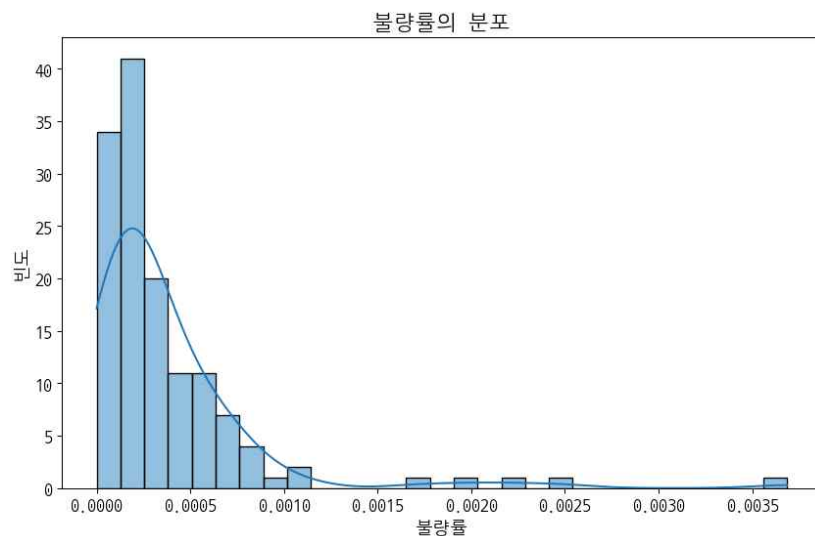
## - 전체 변수간 상관관계 히트맵



비슷한 공정끼리는 양의 상관관계가 존재하는 것으로 파악되었다.

### ○ 종속변수 설정

- 불량률 = 불량수량 / 총 수량



불량률의 분포에 따라 종속변수를 휴리스틱하게 결정하였다. 불량률이 0.001 이상일 경우 불량 공정으로 판단하였고, 136개의 배정번호 중 7개의 배정번호의 레이블을 불량으로 설정하였다.

## ○ 제조 데이터 처리과정

### - 데이터 정제

결측치는 위의 EDA에서 확인한 바에 따라 제거한다. 세부적인 데이터 전처리 사항은 분석사항에 기입하였다.

### - 데이터 품질 지수

인공지능 모델을 학습하기 위해서는 고품질의 데이터가 필요하다. 결측치나 중복 데이터가 존재하거나, 유효범위를 벗어난 값은 모델이 패턴을 학습하는 것을 방해하여 높은 성능을 기대하기 어렵다.

EDA에서 파악한 내용을 바탕으로 데이터 정제 후 한국데이터산업진흥원에서 작성한 『데이터 품질진단 절차 및 기법 v1.0』에서 제시한 기준에 따라 데이터 품질을 만족하였는지 확인하였다.<sup>[4]</sup>

1. 완전성 : 필수항목에 누락이 없어야 한다.

▶ 결측치를 모두 제거하여 완전성을 충족하였다.

2. 유일성 : 데이터 항목은 유일해야 하며 중복되어서는 안된다.

▶ 중복 데이터 개수가 0임을 확인하여 유일성도 충족되었다.

3. 유효성 : 데이터 항목은 정해진 데이터 유효범위 및 도메인을 충족해야 한다.

▶ 도메인에 대한 지식이 없어 각 변수의 수치가 유효 범위 내에 있는지는 파악할 수 없었지만 데이터 분포를 살펴보았을 때 극단적으로 치우친 값은 없다고 판단되어 유효성을 충족하였다고 결론지었다.

	count	mean	min	25%	50%	75%	max	std
TAG_MIN	2935045	2022-04-20 13:38:10.622449408	2022-01-03 11:22:09	2022-03-12 01:07:47	2022-04-22 15:30:31	2022-06-06 01:16:07	2022-07-19 19:08:59	NaN
배정번호	2935045.0	128442.982429	102410.0	119448.0	129889.0	139116.0	148069.0	12640.665011
건조 1존 OP	2935045.0	69.893235	47.2532	68.4293	70.5171	72.3776	87.2995	4.016107
건조 2존 OP	2935045.0	20.440534	0.000119	18.9149	21.2896	23.3827	47.5395	5.21615
건조로 온도 1 Zone	2935045.0	100.00624	97.3421	99.8146	100.002	100.191	102.469	0.435954
건조로 온도 2 Zone	2935045.0	100.020349	97.8706	99.8903	100.019	100.161	101.843	0.361287
세정기	2935045.0	67.719346	60.6244	66.5705	67.6978	68.98	71.4901	1.630345
소입1존 OP	2935045.0	75.643363	0.00085	64.9612	82.2102	95.36815	100.0	25.161569
소입2존 OP	2935045.0	54.860442	8.62001	53.3257	55.6648	57.5709	77.2709	4.427754
소입3존 OP	2935045.0	53.859311	0.043705	52.3886	53.8857	55.4134	66.015	2.66436
소입4존 OP	2935045.0	71.090698	0.006244	69.6794	71.0464	72.4782	87.3907	2.557019
소입로 CP 값	2935045.0	0.448859	0.005096	0.448442	0.450062	0.451706	0.909111	0.018879
소입로 CP 모니터 값	2935045.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
소입로 온도 1 Zone	2935045.0	859.214364	840.298	857.958	859.577	860.259	877.228	3.635172
소입로 온도 2 Zone	2935045.0	860.002469	855.929	859.777	860.022	860.249	866.034	0.556664
소입로 온도 3 Zone	2935045.0	860.002878	858.28	859.829	860.002	860.171	870.119	0.351654
소입로 온도 4 Zone	2935045.0	860.005963	857.992	859.843	860.0	860.158	882.148	0.455103
슬트 컨베이어 온도 1 Zone	2935045.0	283.998666	266.23	274.754	284.591	293.348	298.53	9.514179
슬트 컨베이어 온도 2 Zone	2935045.0	279.925922	266.426	273.498	280.012	286.333	291.696	6.612762
슬트조 온도 1 Zone	2935045.0	331.806225	328.161	331.867	332.017	332.141	332.717	0.782815
슬트조 온도 2 Zone	2935045.0	332.177279	328.073	332.178	332.423	332.626	333.179	0.873378

4. 일관성 : 데이터가 지켜야 할 구조, 값, 표현되는 형태가 일관되게 정의되고, 서로 일치해야 한다.

▶ 측정 시간은 datetime, 배정 번호와 불량 여부는 int, 이를 제외한 모든 변수들의 type은 float 형태로 일관성을 충족하였다.

TAG_MIN	datetime64[ns]
배정번호	int64
건조 1존 OP	float64
건조 2존 OP	float64
건조로 온도 1 Zone	float64
건조로 온도 2 Zone	float64
세정기	float64
소입1존 OP	float64
소입2존 OP	float64
소입3존 OP	float64
소입4존 OP	float64
소입로 CP 값	float64
소입로 CP 모니터 값	float64
소입로 온도 1 Zone	float64
소입로 온도 2 Zone	float64
소입로 온도 3 Zone	float64
소입로 온도 4 Zone	float64
슬트 컨베이어 온도 1 Zone	float64
슬트 컨베이어 온도 2 Zone	float64
슬트조 온도 1 Zone	float64
슬트조 온도 2 Zone	float64
불량여부	int64

5. 정확성 : 변수 간에 상관관계가 있다고 판단되었을 때 측정하는 값이다.

▶ 현재 변수들은 다른 공정일 경우 상관관계가 없는 것으로 파악되어 정확성이 충족되었다고 판단하였다.

\* 데이터 품질을 측정한 코드는 EDA.ipynb에 명시하였다.

## □ 분석모델 개발

### ○ 분석 방법론 및 전략

#### - 분석 방법론

각 배정번호 내의 공정 시간과 분당 생산량이 일정하지 않아 개별 공정 시간대에 대해 라벨링을 진행하는 것이 적합하지 않다고 판단하였다. 또한, 동일한 시간에 기록된 로그가 같은 제품에 대한 것이 아니라고 보인다. 예를 들면, 특정 시간에 기록된 소입로와 세정기의 로그가 같은 제품을 의미하지는 않는다. 따라서, 공정이 완료된 후에 전체 불량률을 바탕으로 라벨링을 진행하며, 이를 통해 공정의 불량 여부를 판단하기로 한다.

배정번호에 해당하는 불량률이 임계값(예: 0.1%) 이상인 경우, 공정 자체를 불량으로 판단하여 해당 배정번호에 속하는 모든 시계열 데이터를 ‘불량’으로 분류한다. 즉, 불량 배정번호에서 일정 구간(예: 60초)만큼 데이터를 잘라내어 그 구간의 시계열 데이터를 전부 ‘불량’으로 라벨링한다. 이렇게 구분된 ‘불량’ 및 ‘정상’의 시계열 데이터를 이용하여 시계열 딥러닝 모델인 LSTM, GRU, TCN, Transformer(encoder)를 이용해 분류(classification)를 진행한다.

딥러닝이 아닌 전통적인 머신러닝 기법들은 변수(feature)를 직접 선택하고 구성하는 과정이 필요하다. 해당 시계열 데이터의 플롯을 살펴본 결과, 시계열의 특성을 시각적으로 파악하기 어려워 직접 특성 엔지니어링을 수행하는 것은 적절하지 않다고 판단하였다. 따라서 스스로 변수의 규칙을 학습하는 시계열 딥러닝 모델을 분석에 사용하기로 한다.

분석 결과 가장 좋은 성능을 보인 모델을 최종 모델로 선정하고, shapley value를 이용하여 모델이 공정을 불량으로 탐지한 경우, 모델이 불량으로 예측을 하는데 영향을 준 변수가 무엇인지 파악하는 요인분석을 진행한다.

## - 해당 아이디어의 필요성

각 공정의 데이터가 일정하지 않고, 동일한 시간에 기록된 로그가 반드시 같은 제품을 의미하지 않는다는 문제점을 고려할 때, 공정 완료 후 전체 불량률을 바탕으로 라벨링하는 방법은 각 제품의 최종 불량 여부를 보다 정확하게 반영할 수 있다. 또한, 배정번호 별로 일정 구간의 시계열 데이터를 '불량' 또는 '정상'으로 구분하여 분류할 수 있어, 분석 과정에서의 복잡성을 줄이고 효율성을 높일 수 있다.

딥러닝 모델들은 데이터로부터 자동으로 종속 변수와 독립 변수 간의 규칙을 추출할 수 있다. 이는 공정 데이터의 복잡성과 다양성을 더 잘 다루면서, 노동 집약적인 특성 엔지니어링 작업을 줄일 수 있다. 특히 시계열 딥러닝 모델은 공정 데이터의 순차적 패턴과 관계를 파악하는 데에 유리하다. 이런 모델의 사용은 불량률 예측의 정확성을 높이는 데에 기여한다.

본 방법론은 공정이 실시간으로 진행됨에 따라 데이터를 실시간으로 분석하고, 설비 이상 여부를 빠르게 판단할 수 있다는 이점이 있다. 이를 통해 생산 효율성 향상과 제품 품질 개선이 가능하다.

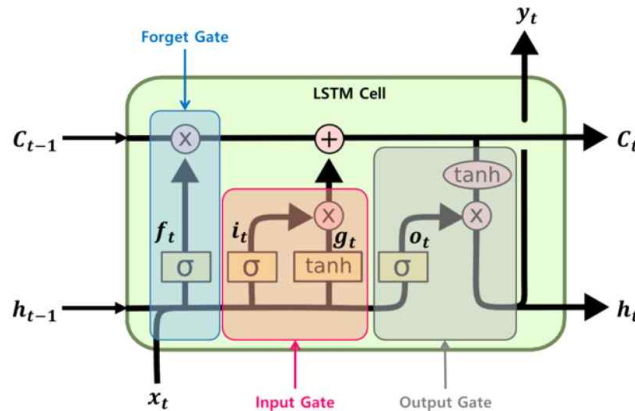
## - 지도학습의 분류(classification) 선택 이유와 중요성

불량률이 임계값을 넘어가지 못하는 공정들을 불량으로 정의했기 때문에 불량공정과 정상공정에 대한 레이블을 정의할 수 있어 분류문제로 접근이 가능하다. 분류 문제는 명확한 카테고리 레이블이 출력값으로 나오기 때문에 쉽게 모델의 예측을 쉽게 이해하고 해석할 수 있다.

또한 정확도(accuracy), F1-score, AUROC 등 다양한 지표를 이용하여 모델의 성능을 객관적이고 다양한 관점에서 평가할 수 있다.

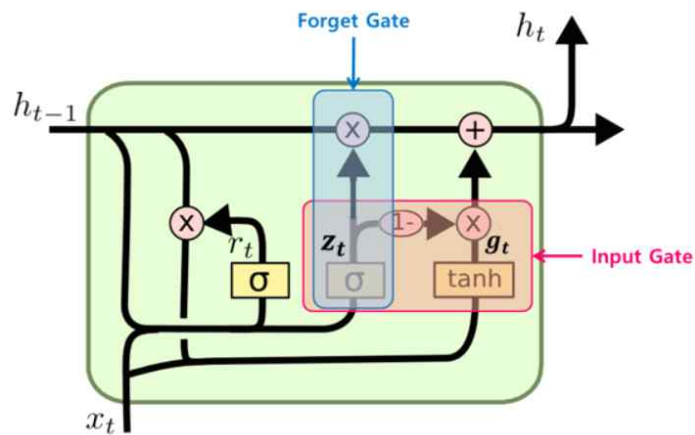
○ 모델 설명

- LSTM ( Long Short Term Memory )



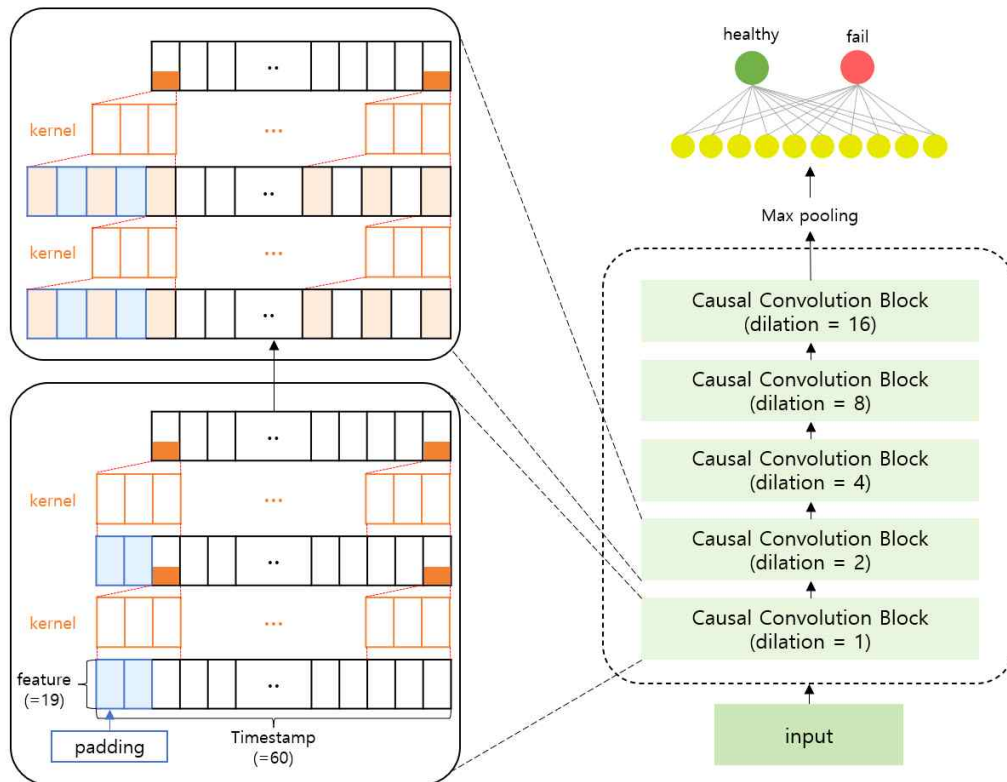
Vanilla RNN의 장기 의존성 문제를 해결한 구조로 단기 상태(short-term state) 와 장기 상태(long-term state)의 hidden vector로 시계열을 학습하는 아키텍처이다. 본 프로젝트에서는 input\_dimension은 입력변수의 개수인 19, hidden vector size를 512, output\_dimension을 2로 LSTM 아키텍처를 구축한다.

- GRU ( Gated Recurrent Unit )



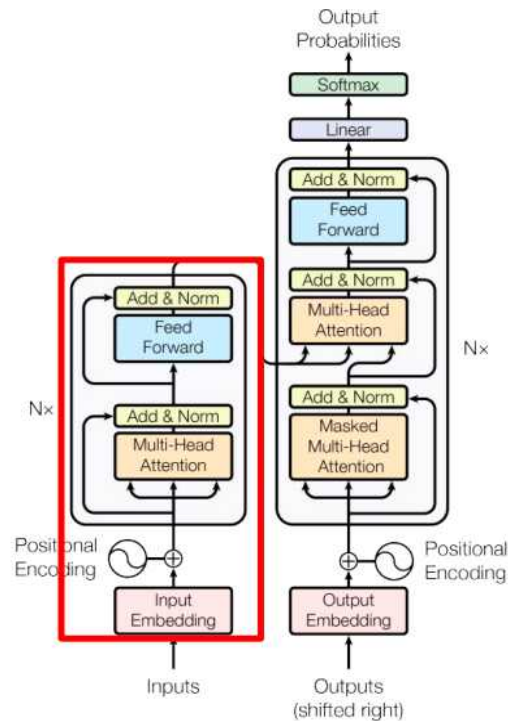
LSTM의 간소화된 버전으로, 게이트 메커니즘을 사용하여 정보 흐름을 조절하면서 장기 의존성을 학습하지만, LSTM보다 더 적은 파라미터를 가지지만 성능은 LSTM만큼 뛰어난 아키텍처이다. LSTM hyperparameter와 동일하게 설정한다.

- TCN ( Temporal Convolutional Network )<sup>[5]</sup>



시계열 데이터를 처리하기 위해 1D 합성곱 레이어와 dilation 메커니즘을 결합한 아키텍처이다. 본 프로젝트에서는 시계열의 시간이 짧은 것을 감안해 구조를 효과적으로 조정한다. 위 그림에서 보듯이 kernel size를 3으로, casual convolution block을 5개로 설정한 뒤 이진분류를 한다.

- Transformer ( encoder )<sup>[6]</sup>



Transformer는 어텐션 메커니즘을 기반으로 순서 정보를 고려하지 않고 데이터 간의 관계를 효과적으로 인식하여 자연어 처리 및 다른 태스크에서 뛰어난 성능을 보이는 아키텍처이다. 그 중 Encoder는 입력 시계열 데이터의 복잡한 패턴과 상호작용을 효과적으로 인코딩하여 고정 길이의 표현으로 압축할 수 있기 때문에, 이를 통해 시계열 분류 작업에 적합한 특징 추출이 가능해 본 프로젝트에 적용한다. 자연어처리 task에서 사용하는 word embedding layer를 거치지 않고 데이터의 feature 개수를 감안해 attention head는 3개로 구성하였다.



## ○ 분석 절차

### - 전체 프로세스



1. 각 배정 번호의 불량률을 구한 뒤 임계치(threshold)보다 불량률이 높은 공정은 불량 공정으로 정의한다. 공정 별로 연속적인 시간대에 대해서 data를 추출한 뒤 불량 공정에 속하면 1로 정상 공정에 속하면 data를 0으로 라벨링한다.
2. 시계열 데이터의 분류가 가능한 모델들에 대해 성능비교를 한 뒤 최고 성능의 모델을 채택한다.
3. 가장 성능이 좋은 모델을 기반으로 shapely value를 추출한 뒤 어떤 변수가 불량에 영향을 주는지 요인 분석한다.

## ○ 데이터 전처리 setting 설정

1. 불량률을 0.001을 기준으로 0과 1로 라벨링 한 이유
  - ▶ 불량률의 분포에 따라 휴리스틱하게 0.001이 optimal point라고 판단한다.
2. null값이 있는 행을 제거 한 이유
  - ▶ end to end model인 시계열 딥러닝 모델은 raw data를 넣어서 학습을 시킨다. 보간값은 실제 값이 아니기 때문에 학습시 왜곡된 영향을 줄 수 있다. 실제 데이터만을 학습한 더 신뢰도 있는 모델을 만들기 위해 결측치를 제거한다. 또한 약 30만개의 데이터에서 컬럼 별 결측치가 평균 300개 이하로 제거하여도 정보 손실이 적을 것이라고 판단한다.
3. 시간대 표현을 60초로 설정한 이유
  - ▶ 120, 180초로 설정 시 label이 1인 time series의 개수가 현저히 줄어든다.

○ 데이터 전처리 세부 과정

1. ‘불량률’ = ( ‘불량수량’ / ‘총수량’ ) \* 100을 정의하고 0.1보다 높으면 1로 낮으면 0으로 배정번호 별로 라벨링한다.
2. data.csv의 결측치가 존재하는 행을 제거한 후 배정번호 별 time series를 추출한다.
3. 연속적인 60초의 시계열을 시간이 겹치지 않게 추출한다. 또한 변수 ‘TAG\_MIN’ , ‘배정번호’ 제거한다.
4. 레이블이 1인 time series는 1,569개이고 레이블이 0인 time series는 34,943개로 레이블의 불균형 때문에 2000, 3000, 4000 ... 개수만큼 임의 추출하여 shuffle한다.
5. shuffle한 data로 train, valid, test 를 6 : 2 : 2 로 분할하고 StandardScaler로 정규화한다.
6. 정규화가 끝난 data들을 tensor로 변환한다.

○ 모델링

제안한 모델의 hyperparameter setting은 다음과 같다.

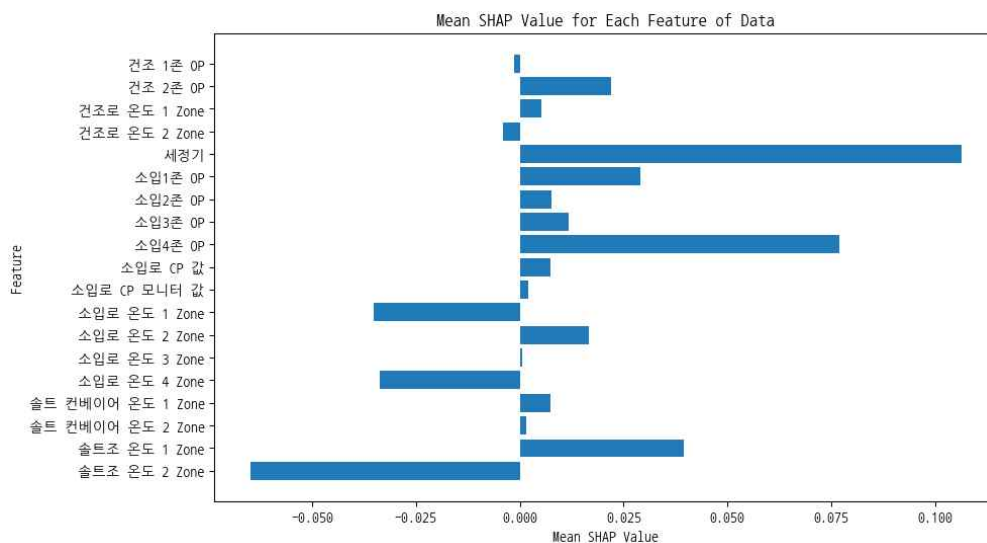
	LSTM	GRU	TCN	Transformer (encoder)
Optimizer	Adam	Adam	Adam	AdamW
Learning rate	0.001	0.001	0.001	0.001
Loss function	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss
Epoch	200	200	200	200
Batch size	256	256	256	256
Hidden size	512	512		
In channels (input feature)	19	19	19	19
Depth			5	
reduced size			10	
Kernel size			3	
Feed forward dim				100
Num layer				12

## ○ 모델 성능 평가

- 본 프로젝트에서 제안한 방법으로 데이터를 추출했을 때  $y=1$ , 즉 불량 공정에 대한 시계열의 데이터 개수는 1569개로 정상 공정에 비해 상대적으로 매우 적은 수를 차지한다. 이러한 데이터 불균형은 모델 학습 시 편향의 원인이 되고 결과적으로 불량품 검출 능력의 저하를 초래할 수 있다.
- 데이터 불균형 문제를 완화하고 정보 손실을 최소화하는 성능을 찾기 위해 양품 데이터에서 무작위로 2000, 3000, 4000, 5000, 6000개의 실험군을 두어 모델을 학습하였다.
- 모델의 성능은 데이터 불균형을 고려하여 macro F1 Score, AUROC(Area Under the ROC Curve), Accuracy(정확도)를 지표로 평가하였다. 모델 성능평가 결과는 결과물 부분에 기재하였다.

## ○ 모델 해석

- SHAP(Shapley Additive exPlanations)[7]은 여러 변수들의 조합 결과에 대해 각 변수의 기여도를 측정하는 방법이다.
- 훈련된 모델에 60초의 time series input을 순차적으로 입력을 할 때 불량( $y=1$ )으로 예측하면 해당 time series에 대해서 Shapley value를 계산한다.
- Shapley value가 아래와 같이 나온다면 건조 2존 op가 다른 변수에 비해 상대적으로 불량으로 분류될 확률에 많이 기여한다고 해석할 수 있다.



## □ 분석결과 및 시사점

### ○ 분석 결과

#### - 모델 분석 결과

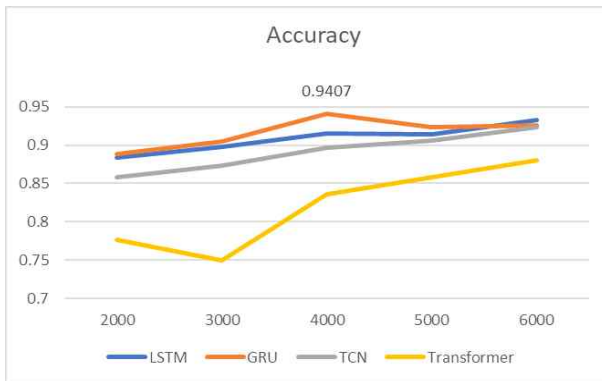
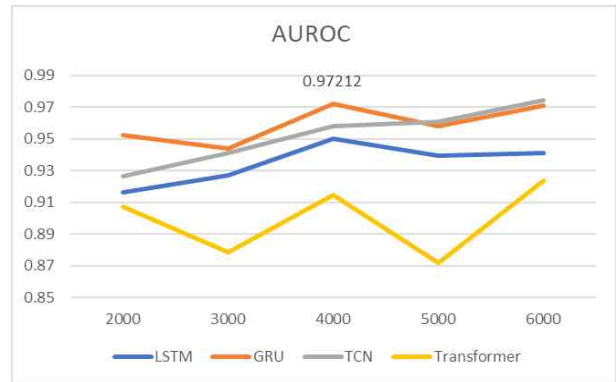
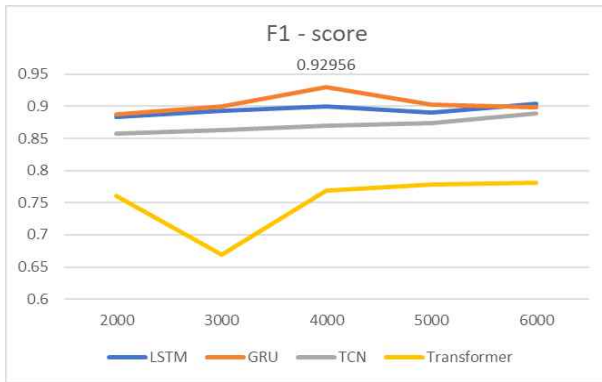
2000	LSTM	GRU	TCN	Transformer
best epoch	104	93	118	96
F1 Score	0.8835	0.88795	0.8578	0.7611
AUROC	0.91616	0.95256	0.92675	0.90728
Accuracy	0.88375	0.88795	0.85854	0.77591

3000	LSTM	GRU	TCN	Transformer
best epoch	108	137	111	100
F1 Score	0.8928	0.90027	0.8628	0.6688
AUROC	0.92722	0.94379	0.9413	0.87841
Accuracy	0.89824	0.90481	0.87308	0.74945

4000	LSTM	GRU	TCN	Transformer
best epoch	123	134	91	95
F1 Score	0.8998	0.92956	0.8697	0.7689
AUROC	0.94989	0.97212	0.9582	0.91457
Accuracy	0.91472	0.9407	0.8968	0.83572

5000	LSTM	GRU	TCN	Transformer
best epoch	168	196	62	197
F1 Score	0.8904	0.9022	0.8733	0.7781
AUROC	0.9397	0.9581	0.9607	0.8716
Accuracy	0.914	0.9231	0.9064	0.8577

6000	LSTM	GRU	TCN	Transformer
best epoch	199	122	140	188
F1 Score	0.9043	0.8983	0.8889	0.7805
AUROC	0.9409	0.9709	0.9741	0.9237
Accuracy	0.9333	0.926	0.924	0.8798



세 가지 지표 모두에서 양품 데이터를 4000개를 추출한 GRU가 가장 좋은 성능을 보인다. 양품 데이터가 증가함에 따라 데이터 정보량의 증가로 성능이 증가하지만 4000개 이후로는 데이터 불균형으로 인해 성능이 떨어짐을 확인하였다.

#### ○ 시사점

- 열처리 제조 공정에서의 실시간 품질 평가는 다양한 기술적, 비용적 제약으로 인해 어렵다. 이는 제조업체들이 공정 중에 발생할 수 있는 문제들을 신속하게 파악하고 조치를 취하는데 큰 장애가 된다.
- 본 프로젝트는 시계열 딥러닝 예측 모델을 이용하여 실시간으로 불량 공정의 시간 표현을 탐지해 요인분석을 진행한다.
- 딥러닝 모델에서 도출된 변수의 중요도(shapley value 기반)를 분석하여, 이를 공정 책임자의 전문 지식과 접목시킬 수 있다. 딥러닝 모델이 제공하는 데이터 기반의 통찰과 책임자의 경험 및 직관을 결합함으로써, 더 정교하고 균형 잡힌 의사결정이 가능하다. 또한, 실시간 데이터 분석을 가능하게 함으로써, 공정 중 발생하는 문제에 신속히 대응이 가능하다.
- 종합적으로, 본 프로젝트의 방법론은 제조업의 품질 관리에 있어서 데

이터 기반 의사결정의 중요성을 강조한다. 특히, 딥러닝 예측 모델을 활용함으로써, 잠재적인 문제에 더욱 신속하고 정확하게 대응할 수 있다. 이는 품질 유지 및 생산 효율 향상에 직접적으로 기여하며, 결과적으로 비용 절감 및 운영 효율성 증진으로 이어질 수 있다는 시사점을 제시한다.

## □ 중소제조기업에 미치는 파급효과

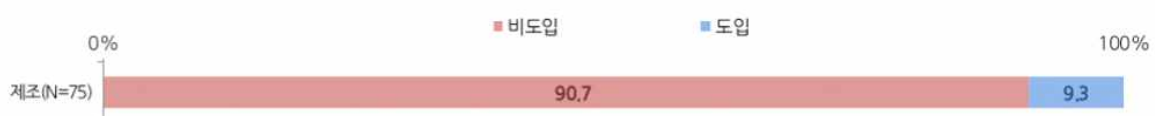
### ○ 현재 중소제조기업의 AI 관련 인식

- 인공지능의 경제적 파급효과가 큰 5대 산업(제조, 교통물류, 금융, 공공안전, 의료)의 종사자수 20인 이상 기업체 368개를 대상으로 한국갤럽과 함께 AI 도입 현황을 조사한 내용 중 제조중소기업에서 응답한 조사결과를 인용하였다.<sup>[8]</sup>
- 조사에 응한 중소기업은 20인 이상 500인 미만은 37개 기업, 500인 이상은 38개 기업으로 확인되었고 그 중 7개의 기업만 AI를 도입하였다. 기업의 종사자 수가 작을수록 AI 도입 비중이 상대적으로 낮은 것으로 나타났다. 중소기업의 경우 AI를 도입하는 데 금전적으로 도입하기 어려울 것이라고 생각된다.

〈표 1〉 주요 산업별 AI 도입 현황 조사 대상

산업	사례 수(개)	
	종사자 수 300인 이상	종사자 수 20인 이상 300인 미만
제조	38	37
교통·물류	36	37
금융	30	43
공공·안전	39	37
의료	병원	22
	의약품 제조기업	10
계	175	193

주: 제조 부문은 종사자 수 500인 이상/미만을 기준으로, 병원 부문은 종합병원 이상/미만을 기준으로 할당 추출



제조중소기업에서 AI 도입 현황

- AI를 적용한 7개의 기업 중 6개의 기업이 AI 도입이 기업에 긍정적인 영향을 미쳤고, 매출 또한 증가했다고 평가하였다.



AI 도입이 경영성과에 미친 영향(%)

〈표 2〉 AI 도입 이후 매출, 인력, 영업비용의 증감(%)

산업	매출			인력			영업비용		
	증가	감소	변화 없음	증가	감소	변화 없음	증가	감소	변화 없음
제조	71.4	0.0	28.6	42.9	28.6	28.6	42.9	14.3	42.9

- AI 기술을 도입하지 않은 기업 중에서도 약 29.4%의 기업이 도입할 계획이 있다고 응답하였다.

〈표 3〉 AI 기술 도입 의향(%)

산업	도입할 계획 있음						도입할 계획 없음	잘 모르겠음
	1년 이내	2~3년 이내	4~5년 이내	6년 이후	도입 시기 미정	계		
제조	1.5	5.9	4.4	2.9	14.7	29.4	36.8	33.8

- 제조 업종에서 품질관리나 이상 징후 탐지를 위해서는 서버를 설치해야 하는데 이러한 도입 비용 또한 높아서 중소기업 입장에서는 기술 도입을 하는데 문제가 있는 것으로 보인다. 또한 제조 업종의 경우, AI로 인해 사고가 발생했을 때 공정 중단이나 불량품이 발생할 뿐 아니라 인재 발생 가능성 또한 존재하기 때문에 이러한 것을 책임질 소재가 불명확한 것이 가장 우려된다고 응답한 것으로 보인다.

〈표 5〉 세부 산업별 AI 기술 도입에 장애가 되는 내부요인(%)

산업	높은 도입 비용	복잡하고 이해하기 힘든 알고리즘	역량을 갖춘 신규인력 채용 어려움	기존 직원의 역량 부족	취약한 기업 내 IT 인프라	내부 가용 데이터 부족	운영 프로세스 개편 비용
제조	42.7	8.0	8.0	13.3	8.0	8.0	5.3

〈표 6〉 세부 산업별 AI 기술 도입 시 우려 사항(%)

산업	AI 시스템이 만든 의사결정 및 행동의 법적 책임	AI의 잘못된 의사결정	AI 사이버보안 취약성	일자리 감소 등 AI 사회적 리스크	AI 실패로 인한 고객신뢰 하락	규제 비준수 리스크
제조	21.3	24.0	21.3	16.0	10.7	4.0



○ 중소제조기업의 AI 도입에 장애되는 요인을 해결한 모델

- 즉, 현재 중소제조기업들은 AI를 도입할 의향은 가지고 있으나, AI 학습 시 필요한 인력, 비용 등의 이유로 도입하기에 어려움을 가지고 있다. 이와 더불어 AI가 잘못된 의사결정을 하였을 때 불량품이 발생하거나, 공정이 중단하는 등의 문제를 우려하는 것으로 확인되었다.
- 본 프로젝트는 이러한 중소기업들의 우려를 해결할 수 있는 모델을 제안하였다. F1-Score는 약 0.93, AUROC는 약 0.97, 정확도는 약 0.94로 AI의 예측 성능을 향상하였으며, 설명 가능한 인공지능을 이용하여 의사결정에 영향을 준 변수들을 확인할 수 있어 모델의 의사결정의 요인을 직접적으로 확인할 수 있다는 장점이 있다.
- 제안한 모델을 사용하면 많은 중소기업들이 제조공정에서 발생하는 설비 공정 이상을 감지하여 생산이 종료되자마자 설비를 바로 수리하여 불량품 최소화 및 생산 최적화를 달성할 수 있어 도입 비용보다도 높은 수익을 낼 가능성이 존재한다.

○ 유사 공정으로의 확장 가능성

	사업체 수 (개)	중견기업 (%)	중소기업 (%)
열처리	1,177	1.4	98.6

- 통계청(2021)에 따르면, 국내 열처리 공정을 진행하고 있는 기업은 총 1,177개로 그 중 98.6%인 1,160개가 중소기업이다.<sup>[9]</sup> 본 프로젝트에서 사용된 데이터와 같은 오스템퍼링 공정을 사용하는 중소기업은 제안한 모델을 적용하여 신속한 불량 공정 탐지가 가능할 것으로 예상된다.

○ 타 분야으로의 확장 가능성

- 제안한 모델은 공정 자체의 설비 상태 데이터를 기반으로 불량 공정을 예측하기 때문에 타 분야이더라도 설비 상태 데이터와 해당 배정 번호의 불량률이 존재한다면 적용할 수 있다.
- 다만 각 공정이 가지고 있는 설비들의 종류와 개수, 센서의 수집 주기 등이 다르므로 모델의 input shape을 어느 정도의 길이로 설정할 것인지, 전처리 시 결측치를 어떻게 처리할 것인지 등을 추가적으로 고려하여 재학습이 필요하다.



## □ Reference

- [1] Bendikiene, Regita, etal. “Influence of austempering temperatures on the microstructure and mechanical properties of austempered ductile cast iron.” Metals 11.6 (2021): 967.
- [2] <https://k3rea.tistory.com/4290594>
- [3] [http://dhheat.com/kr/austempering\\_continuous\\_heat\\_treatment\\_furnace](http://dhheat.com/kr/austempering_continuous_heat_treatment_furnace)
- [4] 한국데이터산업진흥원, 『데이터 품질진단 절차 및 기법(Ver1.0)』
- [5] Franceschi, Jean-Yves, Aymeric Dieuleveut, and Martin Jaggi. “Unsupervised scalable representation learning for multivariate time series.” Advances in neural information processing systems 32 (2019).
- [6] Vaswani, Ashish, etal. “Attention is all you need.” Advances in neural information processing systems 30 (2017).
- [7] Lundberg, Scott M., and Su-In Lee. “A unified approach to interpreting model predictions.” Advances in neural information processing systems 30 (2017).
- [8] 정보통신연구원, 『주요 산업별 인공지능(AI) 도입 현황 및 시사점』  
<https://www.kisdi.re.kr/report/view.do?key=m2101113025339&masterId=4311435&arrMasterId=4311435&artId=600616>
- [9] 통계청, 「전국사업체조사」, 2021, 2023.11.03, 시도·산업·조직형태별 사업체수, 종사자수(’20~ )  
[https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT\\_1K52D02&conn\\_path=](https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1K52D02&conn_path=)