

# 제4회 K-인공지능 제조데이터 분석 경진대회 보고서

프로젝트명	다이캐스팅 결함 제품 분류: 요인 분석 및 모델 비교
팀명	FEB42
내용요약	<p>이 보고서는 다이캐스팅 데이터셋 분석을 통해 얻은 도메인 지식과 분석 결과를 개괄적으로 제공한다. 본 프로젝트는 다이캐스팅 제조 공정 데이터를 중심으로 진행되며, 해당 데이터셋이 프로젝트의 기초를 이룬다. 먼저, 다이캐스팅 공정을 소개하고 주요 생산 문제점을 논의하여 배경을 제공한다. 본 연구의 주요 목표는 중요한 다이캐스팅 제조 속성을 식별하고 이를 활용하여 불량 제품을 분류하는 다양한 머신 러닝(ML) 및 딥 러닝(DL) 모델을 구축하고 비교하는 것이다.</p> <p>Logistic Regression, One-Class SVM, Long Short-Term Memory (LSTM), 맞춤형 Feedforward Neural Network (FFNN) 모델 등 다양한 알고리즘을 평가하여 분류 성능을 비교했다. 이러한 모델들은 동일한 분류 작업에 대한 다양한 접근 방식을 대표하도록 선정되었으며, 각 모델의 구조와 방법론도 간략히 보고서에 설명되어 있다.</p> <p>데이터 전처리 과정에서는 모델링을 위해 데이터셋을 준비하는 일련의 단계를 상세히 수행했다. 모델 성능은 혼동 행렬(confusion matrix), 분류 보고서(classification report), 그리고 최종 매크로 F1 스코어(macro F1-score)를 통해 평가했다. 평가 결과, 맞춤형 FFNN 모델이 가장 높은 성능을 보여 매크로 F1 스코어 89.02%를 기록했다.</p> <p>마지막으로, 이러한 연구 결과의 중소 제조업체에 대한 실질적 함의를 논의하며, 생산 품질 향상을 위한 이와 같은 연구의 중요성을 강조한다.</p>
<p>상기 본인(팀)은 위의 내용과 같이 제4회 K-인공지능 제조데이터 분석 경진대회 결과 보고서를 제출합니다.</p> <p>2024년 11월 06일</p> <p>팀장 : ATALAY EDA (서명) 팀원 : KEBELI BEYZA NUR (서명)</p> <p>한국과학기술원장 귀중</p>	

## □ 문제정의

### ◦ 다이캐스팅 공정 설명

다이캐스팅은 금속 부품을 생산하기 위해 용융 금속을 고압으로 금형(다이) 캐비티에 주입하는 제조 공정이다. 다이캐스팅 공정의 일반적인 단계는 Figure 1에 나와 있다.

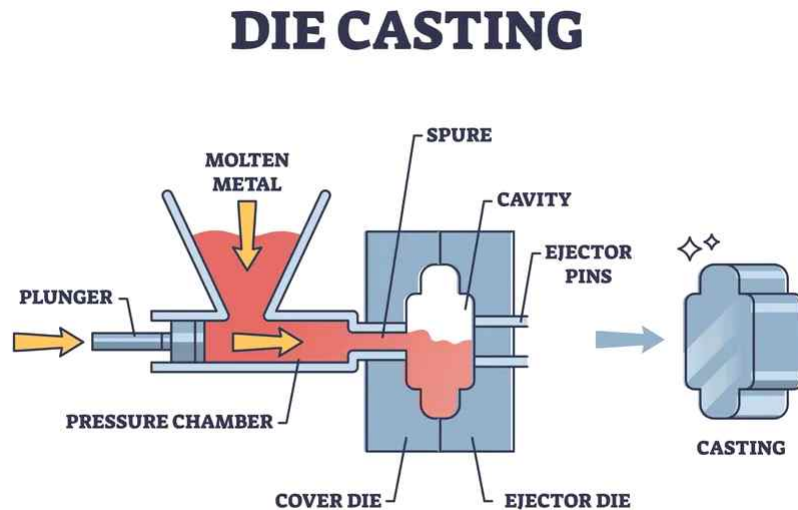


Figure 1: 다이캐스팅 과정

1. 금형 예열: 금형을 예열하여 금속이 원활하게 흐를 수 있게 하고, 주조 품질을 향상시킨다.
2. 용융 금속 주입: 용융된 금속을 고압으로 금형에 정밀하게 주입하여 모든 디테일을 포착한다.
3. 냉각 및 고형화: 금속이 금형 안에서 냉각 및 고형화되어 원하는 형태를 형성한다.
4. 제품 배출: 금형이 열리며 고형화된 제품이 제거된다.
5. 트리밍 및 마감: 여분의 재료를 제거하는 간단한 트리밍 작업을 통해 사용 준비가 완료된 제품을 만든다.

다이캐스팅은 빠른 생산 속도, 정확한 치수, 그리고 매끄러운 표면 마감을 제공하는 여러 가지 장점이 있다. 덕분에 추가적인 가공이나 마감이 필요하지 않은 경우가 많아, 전체 생산 비용을 낮추는 데 도움이 된다. 이러한 이유로, 다이캐스팅은 균일하고 품질이 높은 부품을 대량으로 생산하는 데 매우 적합한 방법이다.

그러나 자동차 엔진, 터빈 블레이드 또는 항공 우주 부품과 같이 복잡한 형상과 정교한 기하학적 구조를 가진 부품을 제조할 때는 대체로 모래 주조, 인베스트먼트 주조, 원심 주조와 같은 다른 주조 방법이 선호된다. 이러한 방법들은 더 세밀한 형태, 복잡한 언더컷, 높은 강도와 정밀도가 요구되는 부품을 만드는 데 더 적합하다.

### - 다이캐스팅의 이슈사항:

다이캐스팅에서 제품이 원하는 형태로 만들어지도록 하는 것이 주요 목표이다. 그러나 여러 문제가 결함이 있는 부품을 생산하게 할 수 있다:

- **불충분한 금형 충전:** 금형 충전이 불충분한 데는 여러 요인이 작용할 수 있으며, 그 중에는 다음이 포함된다:

- **압력 부족:** 다이캐스팅은 용융 금속이 금형을 완전히 채우고 금형 벽에 밀착되도록 높은

압력이 필요하다. 압력이 낮으면 충진이 불완전하게 되어 빈 공간, 약한 부위, 구조적 결함이 있는 부품이 생성될 수 있다.

- **콜드 셋 및 불완전 융합:** 주조 과정에서 금형 내에 사전에 고형화된 “비스킷”이라고 불리는 부분이 유입된 용융 금속과 만나면 콜드 셋이 발생하여 융합이 불완전해질 수 있다. 이는 두 흐름 사이에 생긴 산화층 때문에 단단한 결합이 이루어지지 않아서 발생하며, 제품에 약한 부위를 만든다. 이러한 결함은 부품의 구조적 완전성에 심각한 영향을 미쳐, 압력 하에서 균열이 발생하거나 분리될 수 있다.

- **온도 제어:** 용융 금속과 금형 모두 최적의 온도로 유지되어야 한다. 용융 금속의 온도가 너무 낮으면 조기 고형화되어 충진이 불완전해지고 표면 결함이 발생할 수 있다. 반대로 온도가 너무 높으면 금형의 마모가 심해지고 부품 품질이 떨어질 수 있다. 온도 관리는 변형, 기공, 불균일한 표면 마감과 같은 결함을 방지하기 위해 필수적이다.

- **인적 개입:** 기계 설정 오류나 작업 모니터링의 부족과 같은 인간의 실수도 다이캐스팅 과정에서 결함 제품을 초래할 수 있다.

## ◦ 목표

오늘날 AI(인공지능)는 데이터 활용을 통해 제조업에서 품질을 향상시키고, 비용을 절감하며, 효율성을 개선할 수 있는 중요한 기회를 제공한다. 머신러닝 모델을 활용함으로써 AI는 제조 과정에서 결함이 있는 제품을 예측할 수 있다. 또한 데이터 분석은 결함의 공통 원인을 식별할 수 있게 하여 제조 과정을 최적화하는 데 도움을 준다. 이러한 정보는 결함 제품을 제거하는 데 사용될 수 있으며, 이는 보다 효율적이고 신뢰할 수 있는 생산 시스템으로 이어진다. 최근 보고서에 따르면, AI 기술을 도입함으로써 제조업체는 시장에서 경쟁 우위를 확보하고 끊임없이 변화하는 비즈니스 환경에서 앞서 나갈 수 있다.<sup>[1]</sup>

목표는 결함이 있는 제품을 정확하게 분류하는 모델을 만드는 것이다. 이를 통해 결함에 기여하는 특징을 식별하고 향후 운영을 개선하기 위해 생산 기계의 설정을 최적화할 수 있다. 이 모델을 개발하기 위해서는 제조 데이터에 포함된 정보를 분석해야 하며, 관련 특징의 의미와 값 범위를 파악하고 다양한 모델들을 비교하여 가장 적합한 모델을 선택해야 한다.

## □ 제조데이터 정의 및 처리과정

### ◦ 제조데이터 정의

이 데이터셋은 다이캐스팅 기계와 관련되어 있으며 92,015개의 행을 포함하고 있다. 생산된 제품이 결함이 있는지 여부를 판단하는 것이 중요하다. 첫 번째 단계는 데이터셋에 대한 탐색적 데이터 분석(EDA)을 수행하여 제품 결함에 영향을 미치는 특징을 식별하는 것이다. 데이터셋에는 31개의 열이 포함되어 있지만 아래에 이 분석의 제일 중요한 열들을 설명한다.

변수명	데이터 타입	의미
molten_temp	float64	용탕온도
facility_operation_CycleTime, production_cycletime	int64	설비 작동 사이클시간,제품생산 사이클 시간
low_section_speed, high_section_speed	float64	저속구간속도, 고속구간속도
cast_pressure	float64	주조압력
'biscuit_thickness	float64	비스켓두께
upper_mold_temp1,2,3	float64	상금형온도1,2,3
lower_mold_temp1,2,3	float64	하금형온도1,2,3
physical_strength	float64	형체력
Coolant_temperature	float64	냉각수 온도
sleeve_temperature	float64	슬리브(pressure chamber) 온도
DateTime	datetime	데이터 수집 기간
passorfail	float64	양품불량판정

## ◦ 탐색적 데이터 분석 (EDA)

특징들의 히스토그램 분포를 검토하여 잠재적인 패턴을 확인하였다. Pass 및 Fail 제품에 대한 데이터를 분석한 결과는 데이터의 불균형 문제가 있는 것으로 나타났다. 특히, 불량 제품의 비율이 4.364%임을 확인할 수 있었으며, 데이터의 불균형은 Figure 2에서 확인할 수 있다.

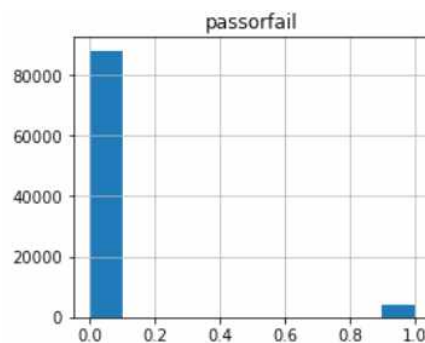


Figure 2: passorfail 분포

원본 데이터셋에서는 “time“과 “date“ 열의 이름이 잘못 지정되어 있었다. “time“ 열에는 날짜가, “date“ 열에는 시, 분, 초 정보가 포함되어 있었다. 먼저 이 열 이름을 올바르게 수정했다. 다음으로, 날짜별 불량품 수를 조사하였다. 불량품이 가장 많이 발생한 날은 2019년 3월 8일로 112개의 결함이 있었고, 가장 적은 불량품이 발생한 날은 2019년 3월 30일로 단 2개의 결함이 있었다.

잠재적인 패턴을 조사하기 위해 이 두 날의 특징에 대한 히스토그램을 작성하였다. Figure 3에서 알 수 있듯이, 불량품이 많은 경우에 “cast\_pressure“ 특성은 대략 100에서 180 사이의 값을 보였다. 불량품이 가장 적었던 날(Figure 4 참조)에는 “cast\_pressure“ 범위가 331에서 334로 이동하였

다. 이는 “cast\_pressure“가 “passorfail“에 상당한 영향을 미칠 수 있음을 시사하였고, 이에 따라 추가 분석을 진행하였다.

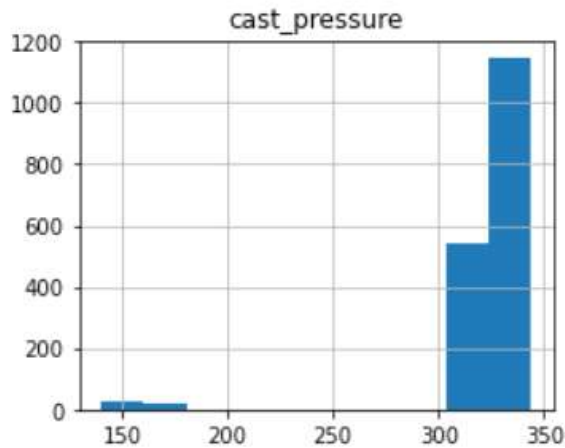


Figure 3: 불량품이 가장 많이 발생한 날

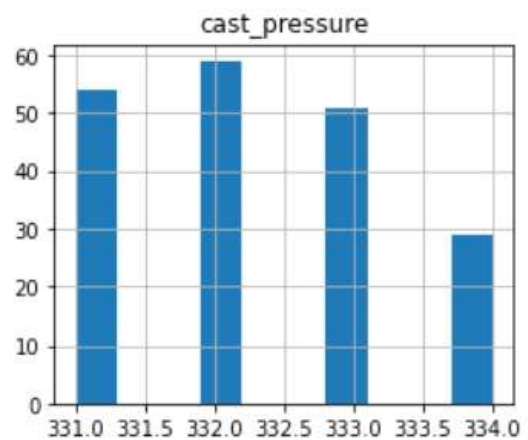


Figure 4: 가장 적은 불량품이 발생한 날

또한, 불량품이 가장 많이 발생한 날의 “biscuit\_thickness“는 대략 0에서 100 사이였고, 불량품이 가장 적었던 날의 범위는 40에서 55 사이였다. 이는 “biscuit\_thickness“의 높은 값이 제품 결함에 기여할 수 있음을 시사한다. 이러한 범위는 Figure 5와 6에 나타나 있다.

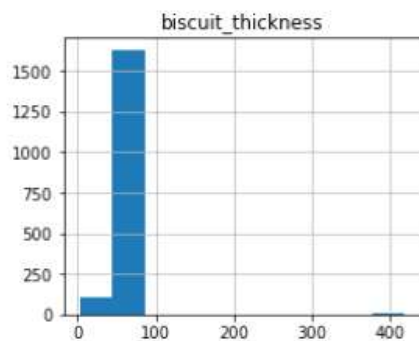


Figure 5: 불량품이 가장 많이 발생한 날

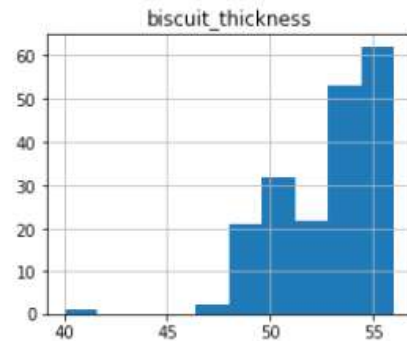


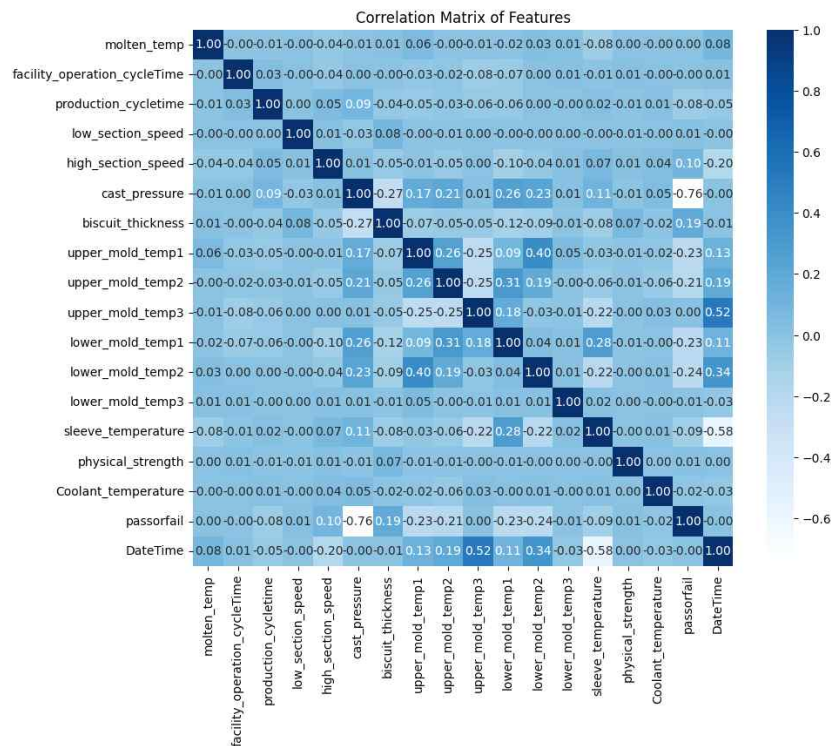
Figure 6: 가장 적은 불량품이 발생한 날

또한, Pass 및 Fail 제품의 모든 수치적 특성에 대한 통계 분석을 별도로 수행하였다. 표 1에서 볼 수 있듯이, “비스킷 두께“와 “캐스트 압력“은 주목할 만한 차이를 보였다:

1. “비스킷 두께“의 경우, 불량품의 최대 값은 60이었고, 합격 제품의 최대 값은 422였다. 이는 “비스킷 두께“가 60을 초과하는 제품은 모두 결함이 있음을 시사한다.
2. “캐스트 압력“의 경우, 합격 제품의 최소 값은 149였고, 불량품의 최소 값은 41이었다. 이는 “캐스트 압력“이 149 미만인 제품은 모두 결함이 있음을 나타낸다.

Product type	Feature	Mean	Min	Max
Fail	Cast pressure	227.784	<b>41.0</b>	339.0
	Biscuit thickness	66.389	0.0	<b>422.0</b>
Pass	Cast pressure	328.970	<b>149.0</b>	348.0
	Biscuit thickness	49.948	1.0	<b>60.0</b>

분석 과정에서 눈에 띄는 특성 중 하나는 “tryshot\_signal“이었다. 이 열은 “D“ 또는 NaN의 값을 포함하고 있다. 다이캐스팅 분야에 대한 연구를 수행한 결과, “D“ 값은 정규 생산 전에 다이캐스팅 기계를 조정하기 위해 사용되는 테스트 런인 데모 샷을 나타낸다는 것을 알게 되었다. 이후 데모 샷이 불량품과 관련이 있는지 확인한 결과, 모든 데모 샷이 결함으로 끝났음을 발견하였다. 총 1,919개의 데모 샷이 있으며, 이는 전체 불량품의 거의 절반에 해당한다. 데모 샷은 이상치로 보이지만, 이용 가능한 불량품의 수가 제한적이므로 데이터를 유지하기로 결정하였다. 마지막으로, 남아 있는 특성들의 상관 행렬(correlation matrix)을 확인한 결과, ‘cast\_pressure’ 변수가 ‘passorfail’ 과 가장 높은 상관 관계를 가지며, 그 값은 -0.76임을 다시 한번 확인할 수 있었다.



## ◦ 데이터 전처리 (Preprocessing)

- 탐색적 데이터 분석(EDA) 과정에서 얻은 결과를 반영하기 위해 “is\_biscuit\_thick” 및 “is\_cast\_pressure\_high” 라는 두 가지 특성을 생성하였다. 이러한 특성은 특정 조건에 따라 0 또는 1의 값을 가진다. 이는 feature engineering을 통해 모델에서 이러한 특성의 중요성을 강조하기 위해 수행된 것이다.
- object 데이터 유형은 분류 문제에 유용한 지표가 아니므로, object 데이터 유형의 특성을 제거하였다.
- ‘date’와 ‘time’ 열을 결합하고 데이터 유형을 datetime으로 변환하여 ‘DateTime’이라는 새 열을 만들었다. 이는 데이터셋의 일관성을 확보하기 위해 수행된 작업이었다.
- ‘date’와 ‘time’을 ‘DateTime’ 특성으로 결합했으므로, 별도의 ‘date’와 ‘time’ 열은 삭제하였다. 불필요한 열을 삭제하는 것은 데이터셋의 유일성을 위해 중요하다. 또한, 각 특성에 있는 null 값의 개수를 확인하였다.

변수명	Null 값 개수
molten_temp	2261
low_section_speed	1
high_section_speed	1
molten_volume	45130
cast_pressure	1
biscuit_thickness	1
upper_mold_temp1	1
upper_mold_temp2	1
upper_mold_temp3	313
lower_mold_temp1	1
lower_mold_temp2	1
lower_mold_temp3	313
sleeve_temperature	1
physical_strength	1
Coolant_temperature	1
passorfail	1

- 하나의 null 값만 있는 열이 동일한 행에 속한다는 것을 발견하여, 해당 행을 삭제하였다. 또한, “molten\_volume” 열은 null 값이 너무 많아 삭제하기로 결정하였다. 그러나 정보 손실을 최소화하기 위해 ‘molten\_temp’, ‘upper\_mold\_temp3’, ‘lower\_mold\_temp3’ 열은 삭제하지 않기로 하였다. 대신, 이들 열의 null 값을 각 열의 평균 값으로 채웠다. 이러한 접근법은 데이터셋이 완전성 요구 사항을 충족하도록 하기 위해 취한 조치이다.

## □ 분석모델 개발

### ◦ 사용한 모델 및 선택 이유:

이 분석 과정에서는 불균형 데이터에서 결합이 있는 제품을 분류하기 위해 로지스틱 회귀(Logistic Regression), 원-클래스 SVM (One-Class SVM), LSTM, 및 피드포워드 신경망(Feed-forward Neural Network) 등 다양한 모델을 선택하여 이러한 데이터가 제기하는 문제를 해결하기 위한 여러 강점과 접근 방식을 활용하고자 했다.

1. **로지스틱 회귀(Logistic Regression)**: 기초적이고 해석 가능한 모델인 로지스틱 회귀는 특성(feature)과 목표 클래스(target class) 간의 관계를 이해하는 데 도움을 준다. 단순하지만, 클래스 가중치 조정을 신중히 하면 불균형 데이터셋에서도 잘 수행될 수 있는 방법론이다.
2. **원-클래스 SVM(One-Class SVM)**: 이상 탐지에 적합한 원-클래스 SVM은 다수 클래스에서만 학습하기 때문에 불균형 데이터셋에서 유용하다. 이 모델은 결함이 있는 제품을 이상치로 식별하려 하며, 소수 클래스의 광범위한 샘플이 필요하지 않은 접근 방식을 제공한다.
3. **LSTM**: 순차 데이터에 특화된 모델인 LSTM은 데이터에서 시간에 의존하는 패턴을 포착하는 데 도움을 준다. 이는 클래스 분류에 중요한 시간적 관계를 드러낼 수 있다.
4. **피드포워드 신경망(Feed-forward Neural Network)**: 데이터의 복잡한 관계를 모델링할 수 있는 피드포워드 신경망은 결함에 기여할 수 있는 미세한 특징 상호작용을 학습할 수 있는 방법론이다. 클래스 가중치 조정이나 재샘플링과 같은 기법을 사용하여 불균형 데이터를 효과적으로 처리할 수도 있다.

이러한 다양한 유형의 모델을 시도함으로써 불균형 데이터셋에서 결함 제품 분류를 위한 가장 강력한 모델을 찾기 위해 간단한 기술과 고급 기술을 모두 탐구하고자 했다.

#### ◦ 분석모델들의 구체적 소개:

##### - 로지스틱 회귀(Logistic Regression)

로지스틱 회귀는 이진 분류 작업에 일반적으로 사용되는 감독 학습 알고리즘이다. 이 알고리즘은 입력 샘플이 특정 클래스에 속할 확률을 예측한다. 로지스틱 회귀는 독립 특성과 이진 목표 변수 간의 관계를 추정하는 데 로지스틱(시그모이드) 함수를 사용하며, 이는 예측된 값을 0과 1 사이의 범위로 매핑한다. 로지스틱 회귀의 개념적 프레임워크는 아래 그림에 나타나 있다.

결함 제품 분류의 맥락에서 로지스틱 회귀는 제품의 특성 값에 따라 결함이 있거나 없는 확률을 평가하는 데 도움을 준다. 주요 장점으로서는 단순성, 해석 가능성, 효율성이 있으며, 이를 통해 효과적인 기준 모델이 된다.

불균형 데이터셋의 경우, 로지스틱 회귀는 클래스 가중치를 조정하여 소수 클래스(결함 제품)에 대한 정확도를 개선할 수 있다.

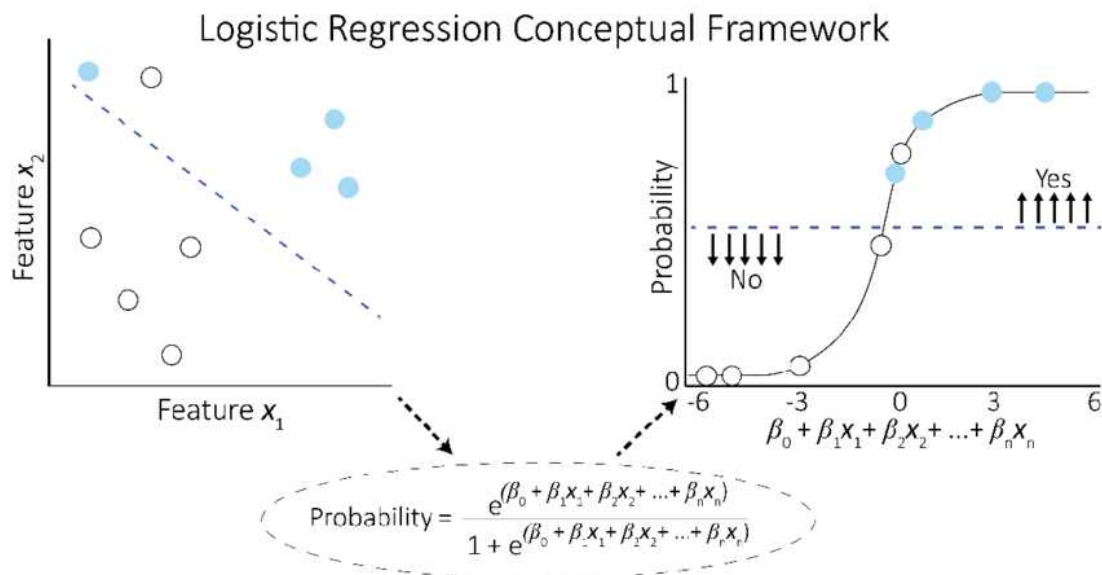


Figure 7: Conceptual Framework for Logistic Regression (Binary Classification) [2]



## - 원-클래스 SVM (One-Class SVM):

원-클래스 SVM(Support Vector Machine)은 이상 탐지를 위해 설계된 비지도 학습 알고리즘이다. 이 알고리즘은 고차원 공간에서 정상 데이터 포인트 주위에 결정 경계를 학습하여, 이를 이상치와 구분한다. 이 특성 덕분에 원-클래스 SVM은 결함이 있는 제품을 탐지하는 경우처럼 드물게 발생하는 이상 패턴이나 사건을 식별하는 데 적합하다. 이상 탐지 작업을 위한 원-클래스 SVM의 표현은 아래 그림에서 확인할 수 있다.

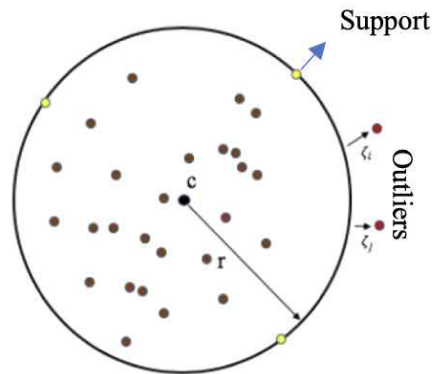


Figure 8: Representation of One-Class SVM Approach for Anomaly Detection [3]

## - LSTM:

LSTM(Long Short Term Memory)은 시퀀스 데이터의 장기 의존성을 처리하도록 설계된 순환 신경망(RNN)의 일종으로, 일반 RNN이 자주 직면하는 기울기 소실 문제를 극복한다. LSTM은 입력 게이트, 망각 게이트, 출력 게이트라는 고유한 셀 구조와 게이팅 메커니즘을 사용하여 정보를 제어한다. 이러한 게이트는 LSTM이 각 단계에서 무엇을 유지하고, 업데이트하며, 버릴지 결정할 수 있게 해준다. 그래서 자연어 처리, 시계열 예측, 순차 데이터 모델링과 같은 작업에 사용하기 효과적인 모델이다.

**목표:** 데이터에서 시간과의 관계가 있는지 확인하고자 한다. LSTM은 이전 정보를 잘 기억할 수 있기 때문에 데이터에 시간적인 패턴이 있다면 LSTM이 이를 학습할 수 있다.

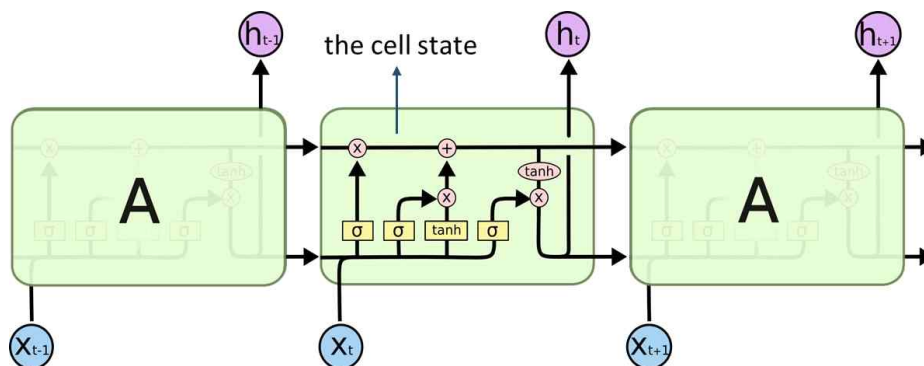


Figure 9: General Look to LSTM [4]

## - Feed-forward Neural Network (피드포워드 신경망):

Feed-forward Neural Network(FFNN)은 데이터가 입력에서 출력으로 한 방향으로 이동하는 기본 신경망이다. 각 층의 뉴런은 가중치가 적용된 입력을 활성화 함수를 통해 전달하여 패턴을 캡처하고 예측을 수행한다. FFNN은 오류를 최소화하기 위해 가중치를 조정하는 역전파(backpropagation) 방식으로 학습하므로 분류와 같은 작업에 적합하다.

### 피드포워드 신경망의 알고리즘:

#### 순전파(Forward Pass):

1. **입력의 가중합:** 각 뉴런은 이전 층으로부터 입력을 받고, 각 입력에 해당하는 가중치를 곱하여 모두 더한다.
2. **활성화 함수를 통한 비선형성 도입:** 가중합은 활성화 함수(예: ReLU 또는 시그모이드)를 통과하여 비선형성을 도입하고, 이를 통해 모델이 복잡한 패턴을 포착하도록 돕는다.
3. **모델의 예측 값:** 모델은 예측 값을 산출한다.

#### 역전파(Backward Pass):

1. **손실 값:** 네트워크의 예측이 실제 목표와 비교되어 예측 오류를 나타내는 손실 값을 생성한다.
2. **그래디언트 계산:** 역전파를 사용하여 모델은 각 파라미터(가중치)에 대한 손실의 그래디언트를 계산한다.

**반복적인 파라미터 업데이트:** 파라미터는 이러한 그래디언트를 기반으로 손실 함수를 최소화하도록 반복적으로 업데이트된다(종종 확률적 경사 하강법과 같은 최적화 방법을 사용하여). 이 과정을 통해 모델의 정확도가 향상된다.

여기서 목표는 피드포워드 신경망을 사용하여 결함이 있는 제품 분류 문제를 개선하는 것이다.

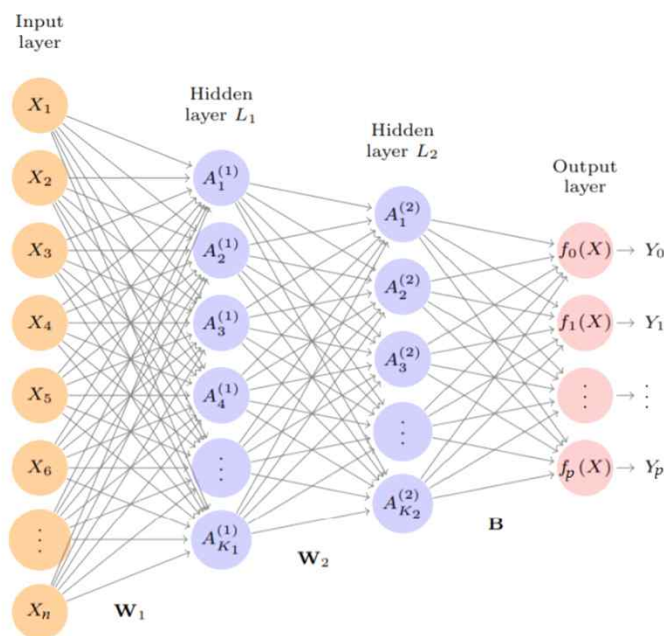


Figure 10: General Look to Feed-forward Neural Networks [5]

## ◦ 분석 모델 개발 과정:

모든 모델에 대해 일관된 컷프로세스를 따랐다: 전처리된 데이터를 가져와 특성 집합(X)과 타겟 변수(y)를 생성했다. 여기서 타겟 변수는 'passorfail' 값을 나타낸다. 특성들이 서로 다른 스케일을 가지고 있기 때문에 스케일러를 사용하여 평균이 0이고 표준 편차가 1이 되도록 표준화했다. 이 표준화는 모든 특성이 비례적으로 기여하도록 한다.

여러 개의 스케일러 중에 Standard Scaler를 선택했다. 이 스케일러는 각 특성의 분산을 데이터셋에 고르게 분포시켜 극단적인 값의 영향을 줄여준다. 이 방법은 특성 스케일의 일관성을 유지하는 데 도움이 되며, 자연적으로 범위가 다른 데이터를 다룰 때 특히 유용하다. 표준화를 통해 모델이 개별 특성의 중요성을 특정 특성의 스케일에 치우치지 않고 해석할 수 있게 한다.

스케일링을 완료한 후, 데이터셋을 60%, 20%, 20% 비율로 훈련, 검증, 테스트 세트로 나눴다. 이후의 프로세스는 각 모델에 따라 약간씩 다르다.

평가를 위해 각 클래스에 대한 기본 분류 지표와 클래스 불균형을 고려한 Macro F1-스코어를 선택했다. 클래스별 지표를 확인하고 전체 정확도 대신 Macro F1-스코어를 사용하여 각 클래스 간의 균형 잡힌 성능을 보장하려고 한다. 또한 각 모델의 confusion matrix를 검토하여 성능을 추가로 평가했다.

### Logistic Regression:

class\_weight 매개변수를 'balanced'로 설정하고 다양한 매개변수 조합을 실험한 후 'liblinear' 솔버를 선택했다. 언급되지 않은 다른 매개변수는 sklearn 라이브러리에서 정의한 기본값으로 설정되었다. 모델 결과는 검증 세트와 테스트 세트 모두에서 평가되었다. 다른 결과는 분석 결과 부분에 나와 있지만, Figure 11의 혼동 행렬(confusion matrix)에서도 확인할 수 있듯이, Pass로 분류되어야 할 제품이 Fail로 예측된 경우가 많다는 점에서 부족함을 보여준다.

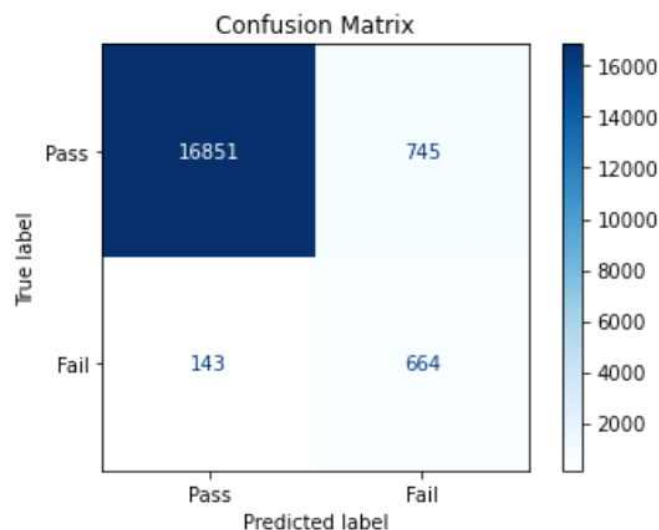


Figure 11: Logistic Regression Confusion Matrix

### One-Class SVM:

원 클래스 SVM을 위해 커널을 'rbf' (방사 기저 함수)로 선택했고 이는 데이터의 비선형 관계를 포

착하는 데 효과적이기 때문이다. 감마를 'scale'로 설정하여 특성 수에 따라 감마 매개변수를 자동으로 조정하고 모델의 복잡성을 최적화하는 데 도움을 준다. null 값은 0.04로 설정했으며, 이는 이상치의 비율에 대한 상한선과 지원 벡터의 비율에 대한 하한선을 나타낸다. 언급되지 않은 다른 매개변수는 sklearn 라이브러리에서 정의한 기본값으로 설정되었다. 모델 결과는 검증 세트와 테스트 세트 모두에서 평가되었다. 여기서도 다른 결과는 분석 결과 부분에 나와 있지만, Figure 12의 혼동 행렬에서도 확인할 수 있듯이, Pass로 분류되어야 할 제품이 Fail로 예측된 경우가 많다는 점에서 부족함을 보여준다.

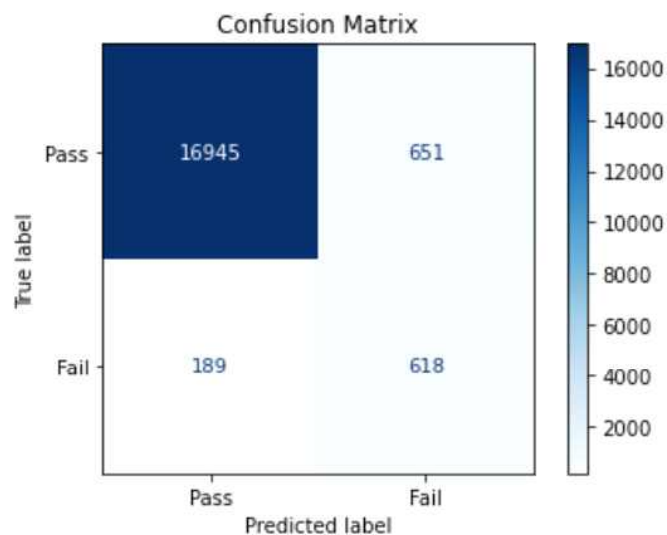


Figure 12: One-Class SVM Confusion Matrix

#### Feed-Forward Neural Network:

모델은 이진 분류를 위한 피드포워드 신경망(FFNN)으로 다음과 같은 구조를 가지고 있다:

1. **입력층:** 64개의 유닛과 ReLU 활성화를 가진 Dense 레이어, 그 뒤에 배치 정규화(Batch Normalization)와 0.3 드롭아웃 비율을 가진 드롭아웃(Dropout) 레이어가 있다.
2. **은닉층:** 16개의 유닛과 ReLU 활성화를 가진 Dense 레이어, 배치 정규화, 그리고 0.3 드롭아웃 비율을 가진 드롭아웃 레이어가 있다.
3. **출력층:** 이진 출력을 위한 단일 유닛과 시그모이드 활성화를 가진 Dense 레이어가 있다.

모델은 Adam 옵티마이저(learning rate: 0.0001)와 이진 교차 엔트로피 손실을 사용하여 컴파일되었다. 조기 중단(Early stopping)을 통해 검증 손실을 모니터링하고 최상의 가중치를 복원한다. 모델은 클래스 불균형을 해결하기 위해 클래스 가중치를 사용하여 최대 50 에폭 동안 배치 크기 128로 학습된다.

모델의 성능은 테스트 세트에 대한 classification report와 macro F1 점수를 사용해 평가한다. 이 결과는 분석 결과 부분에서 더 자세히 다루지만, Figure 13의 혼동 행렬을 통해 FFNN의 우수한 성능을 확인할 수 있다.

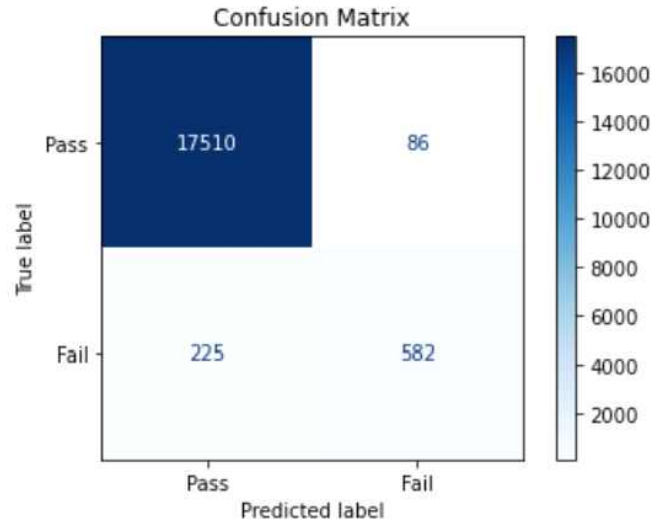


Figure 13: Feed-forward Neural Network Confusion Matrix

#### LSTM:

LSTM 모델은 FFNN과 동일한 구조와 설정을 공유하지만, 데이터의 시간적 의존성을 포착하기 위해 Dense 레이어 대신 LSTM 레이어를 사용한다. 이 모델에서는 타임스텝이 10으로 정의됐다.

모델의 성능은 테스트 세트에 대한 classification report와 macro F1 점수를 사용해 평가한다. 이 결과는 분석 결과 부분에서 더 자세히 다루지만, Figure 14의 혼동 행렬을 통해 LSTM의 성능을 확인할 수 있다. ML 기반 모델들도 더 높은 정확도를 보여줬고 FFNN만큼도 성능이 높지는 않다.

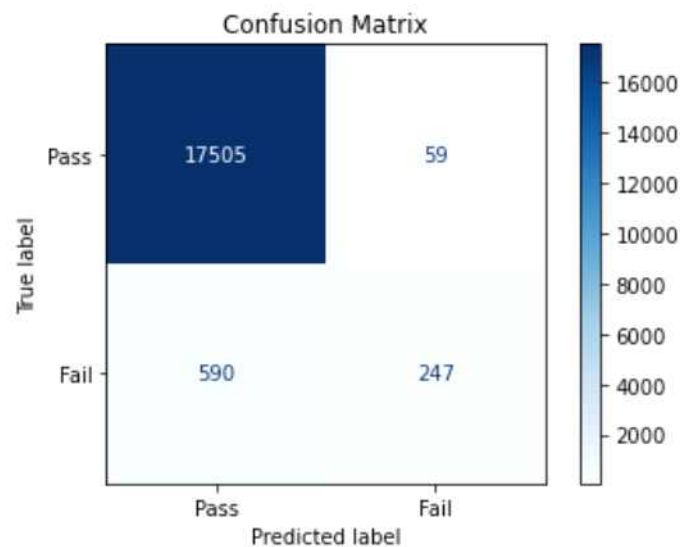


Figure 13: LSTM Confusion Matrix

## □ 분석결과와 시사점

모든 모델의 성능은 아래 표에 요약되어 있다. 모든 값은 테스트 세트에서 모델을 평가하여 얻은 결과이다. 표에서 볼 수 있듯이, 소수 클래스 예측에서도 좋은 성과를 낸 커스텀 FFNN이 가

장 우수한 성능을 보였다.

로지스틱 회귀와 one-class SVM 같은 ML 기반 모델의 성능이 낮은 이유는 타겟 변수와 특성들 간의 복잡한 관계를 포착하는 데 어려움이 있기 때문이라고 판단된다. LSTM 또한 기대보다 낮은 성능을 보였는데, 이는 데이터에 날짜와 시간 정보가 포함되어 있지만 제품의 Pass/Fail 상태가 시간에 따른 패턴을 따르지 않기 때문으로 보인다.

이런 맥락에서 커스텀 FFNN이 이 데이터셋에 가장 적합한 모델로, 순차적 구조에 의존하지 않고도 타겟 변수를 예측하기 위해 피처의 세부적인 정보를 학습할 수 있다고 판단된다.

모든 모델을 개발하는 과정에서 다양한 파라미터 값과 조합을 시도했고, 최종적으로 가장 성능이 좋은 파라미터로 모델을 완성했다. 또한 공정한 비교를 위해 모델 간의 조건을 일관되게 유지했다. 이러한 조건에서 커스텀 FFNN이 가장 적합하고 최적의 모델이라고 생각된다.

Model	Category	Precision	Recall	F1-Score	Macro F1-Score
Logistic Regression	Pass	0.99	0.96	0.97	78.68%
	Fail	0.47	0.82	0.60	
One-Class SVM	Pass	0.99	0.96	0.98	78.56%
	Fail	0.49	0.77	0.60	
LSTM	Pass	0.97	1.00	0.98	70.70%
	Fail	0.81	0.30	0.43	
Custom FFNN	Pass	<u>0.99</u>	<u>1.00</u>	<u>0.99</u>	<u>89.02%</u>
	Fail	<u>0.87</u>	0.72	<u>0.79</u>	

## □ 중소제조기업에 미치는 파급효과

다이캐스팅 공정에서 이 보고서에서 만든 FFNN 아키텍처는 89.02%로 높은 macro F1 score를 기록하여 AI의 예측 성능을 크게 향상시키는 효과를 보여준다. 이러한 성능 향상은 중소 제조업체들이 제품 품질을 개선하고 생산 효율성을 높이는 데 기여할 수 있다.

또한, 설명 가능한 인공지능을 활용함으로써 의사결정의 근거를 명확히 확인할 수 있어 경영진과 운영팀이 데이터 기반으로 보다 정확한 결정을 내리는 데 도움을 준다.

더 나아가, 본 연구에서 사용된 데이터세트와 유사한 다이캐스팅 공정을 사용하는 중소기업들은 제안한 모델을 적용하여 신속하게 불량품을 탐지할 수 있을 것으로 예상된다. 이는 전반적인 제조 품질을 높이고 비용 절감을 이루는 데 중요한 역할을 할 것이다.

## □ REFERENCES

- [1] Abdelaal, M. (2018). *AI in Manufacturing: Market Analysis and Opportunities*. Software AG, Darmstadt, Hessen, Germany.
- [2] Mordensky, S. P., Lipor, J. J., DeAngelo, J., Burns, E. R., & Lindsey, C. R. (2022, February). Predicting geothermal favorability in the Western United States by using machine learning: addressing challenges and developing solutions. In Forty Seventh Workshop on Geothermal Reservoir Engineering.
- [3] Sau, A., Phadikar, S., & Bhakta, I. (2024). Efficient detection of data entry errors in large-scale public health surveys: an unsupervised machine learning approach. *Discover Public Health*, 21(1), 129.
- [4] “An Introduction to Statistical Learning with Application in R” by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.
- [5] “Understanding LSTM Networks” by Christopher Olah.