AI/ML

불량이 적은 공장일수록 AI 만들기가 어렵습니다

방동걸 (주)웨이스 대표이사 2026년 8월 15일

현장에서 가장 자주 듣는 말이 있습니다. "저희는 데이터가 부족해서 AI는 아직 이릅니다." VEXPLOR를 구축하러 간 공장에서 웨이스가 그 말을 들을 때마다 함께 살펴보는 것부터 적어 보겠습니다.

그런데 실제로 부족한 것은 데이터 전체가 아닙니다. 센서값도 생산 이력도 몇 년치가 쌓여 있는 경우가 대부분입니다. 부족한 것은 딱 한 종류, **불량 데이터**입니다. 그리고 이건 공장이 일을 못해서 생기는 문제가 아니라, 일을 잘해서 생기는 문제입니다.



잘 만드는 공장일수록 배울 불량이 없습니다.

1. 잘 만들수록 학습이 어려워지는 구조

불량률이 3%인 공장이 있다고 해봅시다. 제조 현장에서는 좋은 숫자입니다. 그런데 이걸 학습 데이터로 놓고 보면 정상 97에 불량 3입니다. 여기서 결함 유형을 나누기 시작하면 상황이 더 나빠집니다. 자주 나오는 유형 두세 가지가 그 3%의 대부분을 차지하고, 나머지 유형은 1년에 수십 건 수준으로 떨어집니다. 정작 놓치면 큰일 나는 결함일수록 사례가 적습니다.

이 상태에서 흔히 벌어지는 일이 있습니다. 모델을 만들었더니 정확도가 97%로 나옵니다. 훌륭해 보이지만, 아무것도 안 하고 **전부 정상이라고만 답해도** 나오는 숫자입니다. 그래서 이런 문제에서는 정확도를 보면 안 되고, 실제 불량을 얼마나 잡아냈는지(재현율)와 잡았다고 한 것 중 진짜가 얼마인지(정밀도)를 함께 묶은 F1을 봐야 합니다. 저희가 과제마다 성능 지표를 계약 기준으로 못 박을 때 정확도가 아니라 F1을 쓰는 이유가 이것입니다.

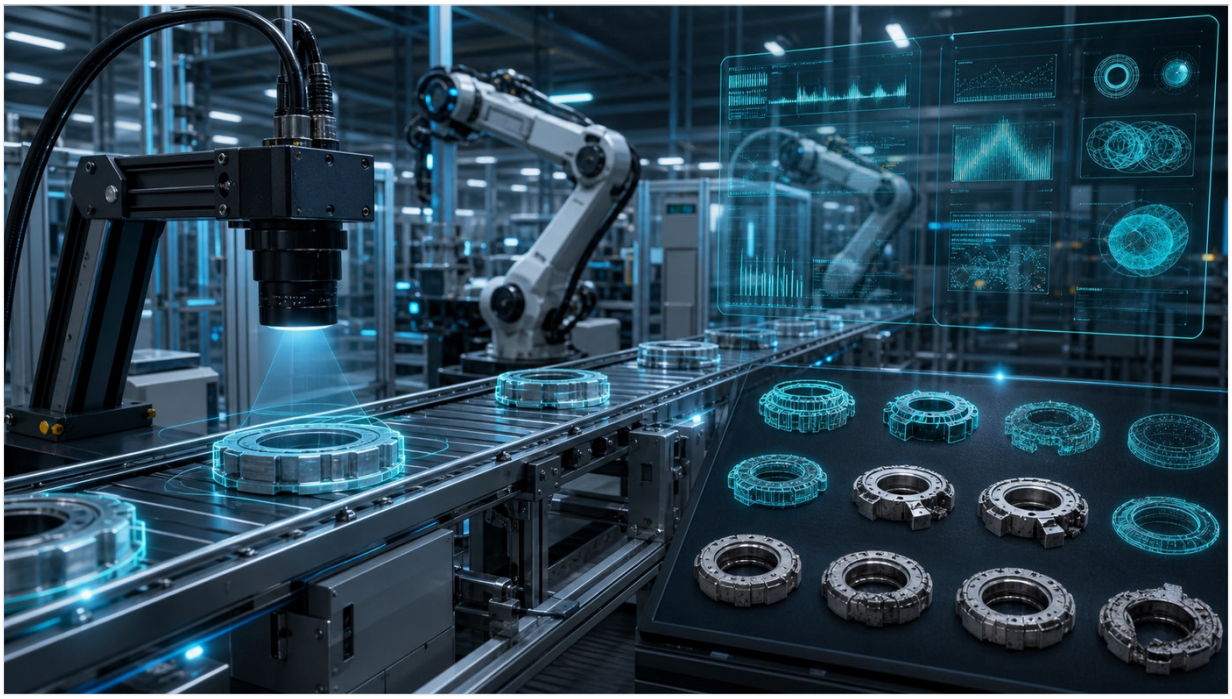
---

2. 불량이 하나도 없어도 시작할 수 있습니다

라벨이 아예 없는 현장이라면 지도학습으로는 출발선에 설 수가 없습니다. 그래서 저희는 반대로 시작합니다. **정상만 학습시키는 방법**입니다.

LSTM Autoencoder는 정상 운전 데이터를 압축했다가 다시 복원하도록 학습합니다. 정상을 충분히 본 모델은 정상을 잘 복원하고, 한 번도 본 적 없는 상태가 들어오면 복원에 실패합니다. 그 복원 오차가 곧 이상 신호입니다. Isolation Forest처럼 성격이 다른 탐지기를 함께 두고 서로 어긋나는 구간을 살피면 오탐도 줄어듭니다.

이 방식의 값어치는 **불량 라벨 0건에서도 착수할 수 있다**는 데 있습니다. 한계도 분명합니다. "지금 평소와 다르다"는 알려주지만 "무슨 불량인지"는 모릅니다. 그래서 이 단계는 끝이 아니라 시작입니다. 이상 구간에 현장이 판정을 붙여 주면 그게 라벨이 되고, 라벨이 모이면 유형을 구분하는 모델로 넘어갑니다. 데이터가 없어서 못 하는 게 아니라, 데이터를 만들면서 가는 겁니다.



학습은 합성으로 해도 평가는 실물로만 합니다.

3. 있는 불량을 더 쓰고, 없는 불량을 만들어 씁니다

라벨이 조금 모이면 다음 문제는 여전히 수가 적다는 것입니다. 여기서 두 가지를 씁니다.

SMOTE는 소수 쪽 사례들 사이를 보간해 중간값에 해당하는 표본을 채웁니다. 단순 복제와 달리 같은 점을 여러 번 찍지 않아 과적합이 덜합니다.

생성 모델은 한 걸음 더 갑니다. 시계열 파형은 TimeGAN으로, 공정 조건표처럼 행과 열로 된 데이터는 CTGAN으로 합성해 학습 표본을 최대 5배까지 늘립니다. 원본 불량이 40건이면 200건 규모로 학습을 돌릴 수 있게 됩니다.

여기에는 반드시 짚어야 할 한계가 있습니다. **합성은 원본에 있는 패턴을 늘릴 뿐, 원본에 없던 결함을 만들어내지는 못합니다.** 아직 한 번도 겪지 않은 고장 형태에 대해서는 아무 도움이 되지 않습니다. 그래서 합성 데이터는 지도학습을 보강하는 수단이지, 2번의 비지도 탐지를 대체하는 수단이 아닙니다. 둘을 같이 두는 이유입니다.

4. 노이즈는 걷어내되 신호는 뭉개지 않습니다

현장 센서 원신호는 그대로 쓰기 어렵습니다. 진동, 전원 잡음, 계측 오류가 섞여 있습니다. 그런데 여기서 흔한 실수가 나옵니다. 노이즈를 없애려고 이동평균을 걸면, 결함의 증거인 **짧고 뾰족한 피크까지 같이 뭉개집니다.** 잡음을 지우면서 찾으려던 것도 지우는 셈입니다.

그래서 **Savitzky-Golay** 평활을 씁니다. 구간마다 다항식을 맞춰 부드럽게 만드는 방식이라, 피크의 높이와 폭을 유지한 채 잡음만 낮춥니다. 이상 신호의 모양이 판정 근거인 공정에서는 이 차이가 결과를 가릅니다.

이상치 처리도 마찬가지입니다. IQR로 범위를 벗어난 값을 걸러내되, 무조건 지우지 않습니다. **제조에서 이상치는 계측 오류일 수도 있지만 바로 그 불량일 수도 있기 때문입니다.** 어느 쪽인지는 통계가 아니라 공정 지식이 판단합니다.

마지막은 특징공학입니다. 용해로 온도 자체보다 승온 속도와 목표 온도 유지 시간이 결함과 더 강하게 연결되는 식으로, 원 신호를 그대로 넣는 것보다 공정이 아는 방식으로 가공해 넣는 편이 훨씬 잘 맞습니다. 데이터가 적을수록 이 작업의 비중이 커집니다. 표본이 적을 때는 모델을 키우는 것보다 변수를 잘 만드는 쪽이 이깁니다.

5. 학습은 합성으로 해도, 평가는 실물로만 합니다

여기가 가장 중요한 대목입니다.

합성 데이터를 검증셋에 섞으면 성능이 잘 나옵니다. 당연합니다. 생성 모델이 만든 불량을 같은 분포에서 온 모델이 맞히는 것이니까요. 그렇게 나온 F1은 **생성기가 만든 문제를 잘 푸는 능력**이지 현장 불량을 잡는 능력이 아닙니다.

그래서 규율은 하나입니다. **합성은 학습에만 쓰고, 검증셋에는 실제 불량만 넣습니다.** 실제 불량이 40건뿐이면 40건으로 k-겹 교차검증을 돌리고 Confusion Matrix를 그대로 열어 봅니다. 숫자가 초라해 보여도 그게 진짜입니다.

어제 저희가 쓴 글에서 소프트웨어 이야기로 같은 결론에 닿았는데, 여기서도 원리는 같습니다. 만드는 쪽과 판정하는 쪽은 분리돼 있어야 합니다. 생성기가 평가까지 겸하는 순간, 그 숫자는 아무것도 보증하지 못합니다.

6. 사실은 그 앞에 라벨 문제가 있습니다

지금까지가 데이터가 적을 때의 기술이라면, 그보다 앞에 있는 문제가 하나 더 있습니다. 라벨이 서로 다른 말을 쓴다는 것입니다.

같은 결함을 어떤 검사자는 크랙이라 적고 어떤 검사자는 균열이라 적습니다. 라인마다 코드 체계가 다르고, 검사 장비의 판정 코드와 품질 시스템의 불량 코드가 따로 놉니다. 이 상태에서 증강을 하면 어긋난 라벨까지 5배로 늘어납니다.

그래서 저희는 모델보다 먼저 이기종 시스템의 코드를 의미 단위로 이어 붙이는 작업을 합니다. 여러 시스템에 흩어진 같은 대상을 같은 것으로 묶어 두면, 그때부터 흩어져 있던 사례가 한 덩어리로 세어집니다. 사례 수가 늘어나서가 아니라 따로 세던 것을 같이 세게 돼서 학습 가능한 규모가 되는 경우가 실제로 있습니다.

07

데이터가 없는 게 아니라 정리가 안 된 것입니다

"AI를 하려면 데이터가 3년치는 쌓여 있어야 한다"는 말은 절반만 맞습니다. 양보다 앞서는 조건이 3가지 있습니다. 정상 데이터가 믿을 만한가, 불량 라벨이 일관된 말을 쓰는가, 그리고 합성으로 늘리더라도 평가만은 실물로 하는 규율이 있는가.

이 3가지가 서 있으면 불량 40건으로도 시작할 수 있습니다. 반대로 이게 없으면 데이터가 10년치 쌓여 있어도 모델은 엉뚱한 것을 배웁니다.

그러니 "저희는 데이터가 부족해서 아직"이라는 진단은 한 번 더 들여다볼 필요가 있습니다. 대개는 없는 게 아니라, 아직 세어지지 않은 상태입니다.

웨이스가 하는 일이 궁금하시면: wace.me

WACE(웨이스)는 공장이 스스로 판단하고 멈추지 않게 하는 제조 AI 운영체제 VEXPLOR를 만드는 회사다.

원문 blog.vexplor.com/post/training-ai-without-defect-data
발행 2026년 8월 15일 · 방동걸 (주)웨이스 대표이사

회사 주식회사 웨이스 (WACE Inc.) · 대표이사 방동걸 · 사업자등록번호 439-87-02852
주소 (18462) 경기 화성시 동탄구 동탄영천로 150, 에이동(A동) 1702-1708호 (영천동, 실리콘앨리동탄)
연락처 전화 070-4730-0629 · contact@wace.me
홈페이지 wace.me · linkedin.com/company/wace-inc

원문 홈페이지 링크드인