

프로젝트 파일 별 간단 요약

wb_data.ipynb

- **핵심 목적:** LSWMD.pkl 전체 웨이퍼를 정리하고 라벨·패턴 유무에 따라 df_withlabel, df_withpattern, df_nonpattern 등 파생 데이터를 만든 뒤 클래스 분포(패턴 8종의 비율)와 대표 웨이퍼 맵 시각화
- **주요 전처리:** waferMapDim으로 해상도 파악, failureType/trainTestLabel을 숫자로 매핑해 이후 모델에서 그대로 쓸 수 있게 준비. 영역 별 결함 밀도와 Radon 변환 기반 통계(평균/표준편차)까지 만들어 패턴별 특징 차이를 확인
- **결론:** 기존 실험 요약 Kaggle 데이터 기반 간단한 분석과 다양한 특징 추출 → multi-type feature 조합 후 One-vs-One SVM 학습 → test weighted F1 \approx 72% 다른 특징 추출/모델을 탐색하며 성능 향상을 모색
- **참고:** <https://aiegg-travel.tistory.com/entry/Kaggle-WM-Dataset%EC%9C%BC%EB%A1%9C-Wafer-Map-Defect-Pattern-%EB%B6%84%EB%A5%98-%EC%8B%A4%EC%8A%B5> 해당 블로그를 그래도 사용했으며 기본적으로 널리 알려진 품이 많은 듯 했음
- 불균형 클래스에서는 F1-score를 더 중요 (따라서 F1-score로 모델을 비교 분석)

cnn_test.ipynb

- **모델 구성:** 26×26 이미지로 줄여 단순 3-블록 CNN(16/32/64채널 + FC 128)로 분류 - 훈련,테스트8:2로 분할
- **성능:** test weighted F1 \approx 85% 달성
- **분석:** 기존에 참조 했던 방식보다 성능이 좋은 것으로 보임

TS_model.ipynb

- **아이디어:** 대학원 동기가 먼저 취업하면서 실무 팁을 알려줬는데, 현업에서는 모델 경량화와 빠른 추론 속도가 중요하다고 해서 공부하던 중 찾은 방법론을 적용
- **모델 구성** 동일 데이터를 56×56 (이미지를 좀 키워서 성능 평가함)으로 만들고 Teacher(1→32→64→128)와 경량 Student(1→24→48→96) 구조를 설계한 뒤, Teacher를 먼저 학습시켜 Student에 지식 증류(KLDiv + CE, T=4, $\alpha=0.7$) 진행
- **Teacher 결과:** Weighted F1 \approx 0.84
- **Student 결과:** Weighted F1 \approx 0.61
- **요약:** Teacher는 충분히 학습했으나 Student가 Teacher 분포를 따라가지 못하면서 전체 파이프라인 성능이 오히려 단순 CNN보다 낮음. 중간 피쳐 매칭이나 temperature 재조정이 → Student가 Teacher의 soft 정보를 충분히 흡수하지 못한 상태이므로 α /temperature를 조정하거나 Student 용량을 조금 늘리는 방법을 생각

AE.ipynb

- **모델 개요:** 동일 데이터를 56×56 단순 2-블록 컨볼루션 오토인코더(인코더 3→16→32 채널, 디코더 ConvTranspose)로 재구성 학습
- **결과 해석:** AutoEncoder(AE)로 이미지를 재구성한 뒤, 인코더 잠재벡터 z를 CNN 입력으로 사용했으나 성능이 기대 이하 "잠재벡터를 다시 CNN 입력으로 쓰면 더 좋을까?"라는 가설을 검증했지만 오히려 기존 CNN보다 낮은 성능 - 잠재벡터 z만 사용하는 대신 원본 이미지 vs 재구성 이미지 차이(Residual map)를 추가 입력/증강으로 활용하면 결함 위치 정보를 더 잘 드러낼 수 있을 것이라는 아이디어 도출
- **한계:** "잠재 벡터 z를 CNN 입력으로 사용하면 분류 성능이 향상될까?"라는 단순한 질문에서 시작했지만, 결과는 기대에 미치지 못했음