Journal of the Korean Institute of Industrial Engineers
Vol. 48, No. 6, pp. 615-625, December 2022.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2022.48.6.615

머신 러닝 모델을 활용한 웨이퍼 불량 탐지 및 테스트 항목 효율화

김호영[1,2] · 강필성[2†]

[1]삼성전자 메모리사업부 / [2]고려대학교 산업경영공학부

Machine Learning Model-based Faulty Wafer Classification and Test Item Reduction

Hoyeong Kim[1,2] · Pilsung Kang[2]

[1]Memory Division, Samsung Electronics / [2]School of Industrial & Management Engineering, Korea University

Wafer testing is one of the key components of the semiconductor manufacturing process and aims to balance maximum production with the highest quality. However, there is a problem that it is difficult to preemptively respond to the changing environment due to the ultra-fine semiconductor process, the quality risk caused by the production of various products, and the lack of professional engineers. Therefore, in this work, we present a framework for determining the optimal set of wafer test items representing high defect wafer detection rates using machine learning models. The proposed framework applies an effective sampling methodology to solve category imbalances, mostly composed of good chips, and uses ensemble classification models and important feature selection methods to achieve high classification performance in a short time without direct wafer evaluation. We demonstrate the proposed methodology has a meaningful effectiveness on time reduction through classification accuracy and test item reduction using real DRAM chip datasets.

Keywords: Ensemble Classification, Probe Test, Imbalanced Data, Feature Reduction

1. 서 론

4차 산업혁명으로 인한 산업 구조의 변화로 인공지능(AI), 사물인터넷(IoT), 빅데이터(Big Data) 등의 첨단정보통신 기술이 대두됨에 따라 반도체 수요가 급격히 증가하고 있다(Kim and Seo, 2021). 이에 따라 전세계 반도체 기업들은 글로벌 시장을 주도하기 위해 기술혁신과 기반 시설에 천문학적인 투자를 진행하고 있으며, 이는 결국 고품질 반도체의 최대 생산을 기반으로 한다. 특히 한국은 반도체 클러스터 착공, 대규모 신규 생산 라인 건설 그리고 EUV와 같은 최첨단 설비 확충과 고급 인력 확보를 통해 2020년 세계 반도체 점유율 18.4%로 미국(50.8%)에 이어 2위를 차지하였다(KSIA, 2021). 특히 메모리반도체인 Dynamic Random Access Memory(DRAM)은 시장 점유율 70% 이상을 차지하며 세계 1위를 유지하고 있다. 하지만 365일 24시간 생산 라인이 운영되는 반도체 산업의 특성상 수요 감소로 인한 가격 하락과 재고증가의 리스크가 크기 때문에(Suh and Chung, 2019) 이를 최소화하기 위한 전략으로서 반도체 웨이퍼 테스트 항목 효율화와 시간 단축이 많은 관심을 받고 있다(Hong and Ahn, 2019).

반도체 제조 프로세스는 크게 Fabrication(FAB) 공정, Electrical

이 논문은 2022년도 정부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2022R1A2C2005455)의 성과물이며, 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2021-0-00471, 모델링 & 최적화 기반 오류-free 정보인프라 자율제어 기술 개발).
† 연락저자 : 강필성 교수, 02841, 서울시 성북구 안암로 145 고려대학교 산업경영공학부, Tel: 02-3290-3383, Fax: 02-929-5888,
E-mail: pilsung_kang@korea.ac.kr
2022년 10월 26일 접수; 2022년 11월 22일 수정본 접수; 2022년 11월 23일 게재 확정.

Die Sorting 검사(EDS test), 패키지 검사(package test)의 3개 단계로 구분할 수 있다. 반도체 공정의 대부분을 차지하는 FAB은 반도체 품질과 수율에 가장 큰 영향을 미치는 과정이며(Baek and Han, 2003), 웨이퍼(wafer)는 FAB에서 200개가 넘는 초미세화 단위 공정들을 거치며 수천 개의 칩을 생성한다. EDS test는 프로브 검사(probe test)라고도 불리며(Lee *et al.*, 2000), FAB 공정을 거쳐 웨이퍼 위에 형성된 칩(chip)에 전류를 흘려 불량을 선 발현시켜 양/불을 판별하는 단계이다. 마지막으로 package test 단계에서는 양품 상태의 칩을 잘라(sawing) 조립(assembly)하여 외부 환경으로부터 칩을 보호해주며, 사용자 환경에서 최종 검사를 진행한다(Baek and Nam, 2002). 본 연구에서 다룰 EDS test는 웨이퍼 테스트라고도 부르며, <Figure 1>과 같이 온도에 따라 Pre-Laser Hot(PLH), Pre-Laser Cold(PLC), EDS final hot(EFH) 3가지 단계로 구분된다. 또한 마지막 단계인 EFH는 레이저 리페어(laser repair) 와 final test로 세분화된다(Lee *et al.*, 2009). 각 단계는 목적에 따라 발현시키고자 하는 불량의 유형이 다르며, 단계별 웨이퍼 테스트 프로그램에는 이를 위한 수십 개의 아이템들이 존재한다.

각 온도 단계별로 테스트 아이템이 진행되면서 칩 별 결점 수(Fail Bit) 데이터가 추출되는데, 결점 수를 여유 셀(spare cell)의 집합인 리던던시(redundancy)로 레이저 리페어 하여 대체할 수 있다면 정상으로, 불량 셀이 여유 셀보다 많아 리페어 불가능한 경우 불량으로 구분하게 된다(Shu and Lee, 2002). 이는 정상 칩도 결함이 존재함을 의미하며, 리페어 알고리즘(repair algorithm)을 통해 칩을 어느 수준에서 양/불을 구분할 것인지 결정한다(Smith *et al.*, 1981). <Figure 2>에 도식화한 웨이퍼와 칩의 구조를 살펴보면, DRAM 웨이퍼는 수천 개의 칩들로 이루어져 있으며, 하나의 칩에는 기가 단위의 셀(cell)이 존재한다. 따라서 무결점의 칩은 존재할 수 없으며, 웨이퍼 테스트를 통해 결점 수를 선 발현하여 리페어 한 후 정상 칩으로 분류하는 것이 매우 중요한 과정이라 할 수 있다(Park and Kim, 2015).

반도체 웨이퍼 테스트 항목 효율화와 시간 단축 프로세스는 <Figure 3>과 같다. 앞서 언급했듯이 EDS test는 온도와 발현시

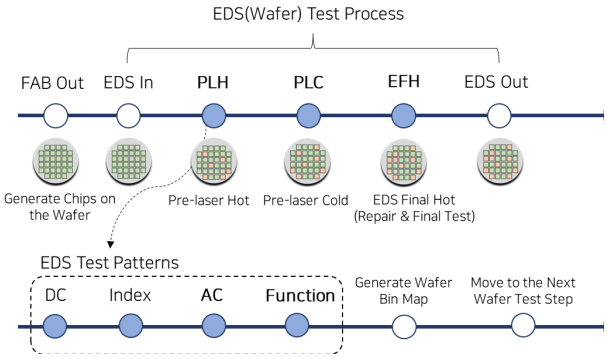


Figure 1. EDS Test Process

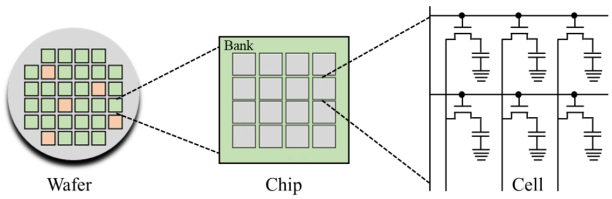


Figure 2. DRAM Architecture

키고자 하는 불량 목적에 따라 세 가지 스텝으로 나뉘며 각 단계별 웨이퍼 테스트 프로그램에 대해 효율화와 시간 단축 작업이 진행된다. 현업에서 최적화가 완료된 제품의 경우 각 스텝별로 15~25분 내외로 테스트가 진행되며, 이는 제품 시작단계 기준 절반 이상 줄어든 시간이다. 아이템 효율화를 위해서는 기존 양산 프로그램에서 추출한 데이터를 기준 데이터(reference data)로 삼고, 중요도가 낮은 항목을 삭제한 후 기준 데이터와 유사한 결점 수와 칩의 양/불량 정보를 가지도록 평가한다. 이는 불량을 선 발현시키는 테스트 아이템 일부를 삭제하더라도 동일한 품질을 유지하기 위함이다. 일반적으로 해당 과정에서 100매 이상의 웨이퍼에 대해 평가를 진행하기 때문에 10일 전후의 평가 시간이 소모되고, 엔지니어의 경험과 역량에 따라 평가 기간과 결과에 편차가 발생한다.

본 연구는 앞서 언급된 문제점(긴 평가 시간 및 엔지니어의 경험과 역량에 따른 평가 기간과 결과의 편차 존재)을 해결하

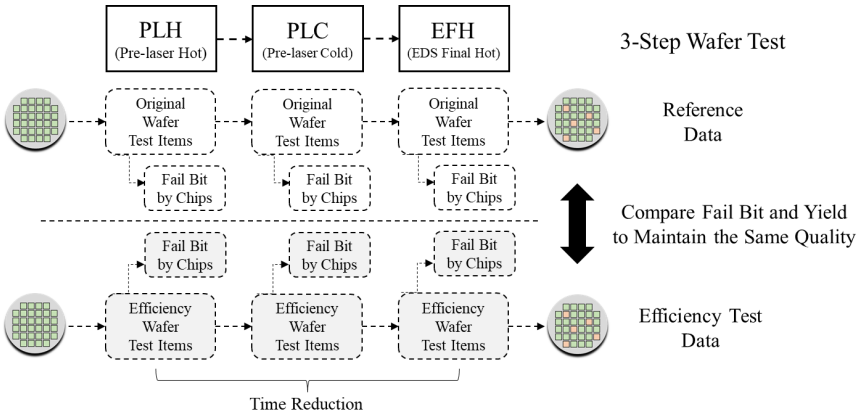


Figure 3. Test Item Efficiency and Time Reduction Process

기 위하여 머신 러닝 기반의 웨이퍼 불량 탐지 및 테스트 항목 효율화 프레임워크(framework)를 제안한다. 제안하는 프레임워크에서는 앙상블 모델을 기반으로 DRAM 웨이퍼 칩에 대한 양/불을 분류하였으며, 샘플링(sampling) 방법론을 활용하여 범주 불균형 문제를 해결하였다(He and Garcia, 2009). 또한 현업 설문조사를 통해 혼동 행렬(confusion matrix)의 양/불량 칩의 정분류, 오분류 비율(False Negative, False Positive)에 대한 중요도를 판별하여 보다 현실적인 평가 지표를 사용하여 성능을 평가하였다. 마지막으로 테스트 항목 효율화와 시간 단축 프로세스를 위해 변수 추출법을 이용하여 중요도가 낮은 테스트 항목들을 선별적으로 제거함으로써 분류 정확도를 최대한 유지하는 효율적인 테스트 항목 집합을 선정하였다.

본 논문에서는 구성은 다음과 같다. 먼저 제2장에서는 반도체 웨이퍼 이미지 분류와 결점 수 데이터를 이용한 패키지 칩 품질 예측 연구 대해 소개하며 차이점을 짚어본다. 제3장에서는 본 연구에서 제안하는 방법론의 전체적인 프레임워크에 대해 언급하고, 제4장에서는 실험에 이용한 데이터셋과 앙상블 모델, 샘플링 방법론, 그리고 변형된 평가지표에 대해 설명한다. 그리고 제5장에서는 실험 결과를 보이고, 마지막으로 제6장에서는 연구의 결론과 한계점, 그리고 향후 연구방향에 대해 간략하게 요약하였다.

2. 관련 연구

반도체 웨이퍼 테스트 데이터와 관련된 대부분의 논문은 WM-811K 웨이퍼 맵(wafer map) 데이터셋을 이용하여 불량 웨이퍼를 탐지하거나, 한 단계 더 나아가 불량의 종류를 구분하는 과업을 수행하였다. 대표적으로 Kim *et al.*(2019)의 연구에서는 Word2Vec을 참고하여 1차원의 스칼라 빈 코드(scalar bin code)를 3차원의 임베딩 벡터(embedding vector)로 변환하여 웨이퍼 맵을 분류하였다. 또한, Jin *et al.*(2019)은 밀도 기반 클러스터링(DBSCAN)을 기반으로 웨이퍼 맵의 불량 패턴을 검출하고 분류하는 연구를 수행하였다. 해당 연구는 outlier 검출과 클러스터링 추출이 동시에 가능하다는 장점을 가지며, 임의의 모양을 가진 불량 패턴을 검출하는 것도 가능하다는 것을 보여주었다. Kahng and Kim(2020)은 자기지도학습(self-supervised learning)으로 사전학습(pre-training)하여 속성 표현(feature representation을 생성해내고, 지도학습(supervised learning)을 통해 미세 소정(fine tuning)을 진행하는 방법론을 제안하였다. 다만, 해당 방법론의 경우 웨이퍼 맵에 crop, cut-out, noise 등과 같은 데이터 증강(data augmentation)기법을 적용하기에 현장에서 활용하기에 다소 무리가 있다는 단점이 있으나, 공정이 미세화 될수록 언제나 새로운 불량이 발생하고 있는 반도체 특성상 자기지도학습을 통한 접근법은 상당한 의미가 있다고 할 수 있다.

반면에 결점 수 데이터셋은 실제 현업에서 반출해야 하는 데이터이기 때문에 이를 활용한 연구가 활발하게 진행되지는 못하고 있는 실정이다. 본 연구는 결점 수를 이용해 테스트 항목을 효율화 하는 것을 목적으로 하며, 본 연구와 동일한 과업을 수행한 연구는 존재하지 않으나, 다음과 같이 유사한 연구를 찾아볼 수 있다. 먼저 regularized singular value decomposition(RSVD)을 통해 특성을 추출하고, k-NN 분류기를 통해 결점 수 맵(fail bit map)의 single bit과 multi bit의 분류를 제안한 연구가 가장 대표적이다(Kim *et al.*, 2015). 해당 논문에서는 정형 데이터인 결점 수 데이터셋이 아닌 칩 별 결점 수를 맵으로 형성한 이미지를 이용하여 결점 수의 형태를 분류하였다. 하지만 실험에 사용한 이미지 개수가 한정적이기 때문에 개념적인 방법론을 제안했다는 수준 이상의 실증적 효과를 확인하지 못했다는 한계점이 있다. Park and Kim(2015)에서는 결점 수 데이터셋을 이용하여 패키지 칩 품질 예측 방법론을 제안하였다. 해당 연구에서는 결점 수 데이터셋의 불균형을 해결하기 위해 SMOTE 방법론을 활용했고, 로지스틱 회귀 모델을 통해 불량을 예측하였다. 해당 연구는 반도체 프로세스의 최종 단계인 패키지 칩에 대한 불량을 예측한다는 점에서 의미를 가지지만, 민감도에만 중요도를 두었기 때문에 특이도가 낮아 일반적으로 적용하기 힘들다는 문제점이 존재한다. Jang *et al.*(2017)은 패키지 테스트 단계에서 반도체 수분 흡습에 따른 불량 발생 정도 평가(preconditioning test) 데이터를 통해 최종 양/불량 예측을 하고자 했다. 앞선 연구와 마찬가지로 해당 논문에서도 양/불량 칩의 데이터 불균형을 해소하기 위해 SMOTE 기법을 사용했으며, Decision Tree, kNN, SVM, Random Forest 등의 알고리즘을 이용하여 분류를 진행했다. 하지만 총 2,892개의 작은 규모의 데이터셋을 기반으로 실험이 진행되었고, 저자가 결론에서 주장하는 추가적으로 진행되는 테스트에 관한 정보가 서술되어 있지 않기 때문에 생산성 향상과 비용 절감 효과를 알 수 없다.

앞선 연구들과 달리 본 논문에서는 현업 설문조사를 통해 양/불량 칩의 정분류와 오분류 비율에 대한 중요도를 재정의하였기 때문에 도메인 지식(domain knowledge) 관점에서 합리적이라 판단할 수 있다. 그리고 불균형 데이터의 해소를 위해 over-sampling 뿐만 아니라 비용민감 학습(cost-sensitive learning) 및 두 가지 방식의 결합도 함께 시도하였으며, 다양한 앙상블 분류 모델을 적용하여 예측 성능의 향상을 추구하였다. 또한, 테스트 항목 효율화와 시간 단축이라는 주제를 통해 현업에서 사용될 수 있는 프레임워크를 제안했다는 점에서 의의를 갖는다.

3. 제안 방법론

3.1 웨이퍼 불량 탐지 및 테스트 항목 효율화 프레임워크

제안하는 머신 러닝 기반 웨이퍼 불량 탐지 및 테스트 항목 효율화를 통한 시간 단축 연구의 프레임워크는 <Figure 4>와

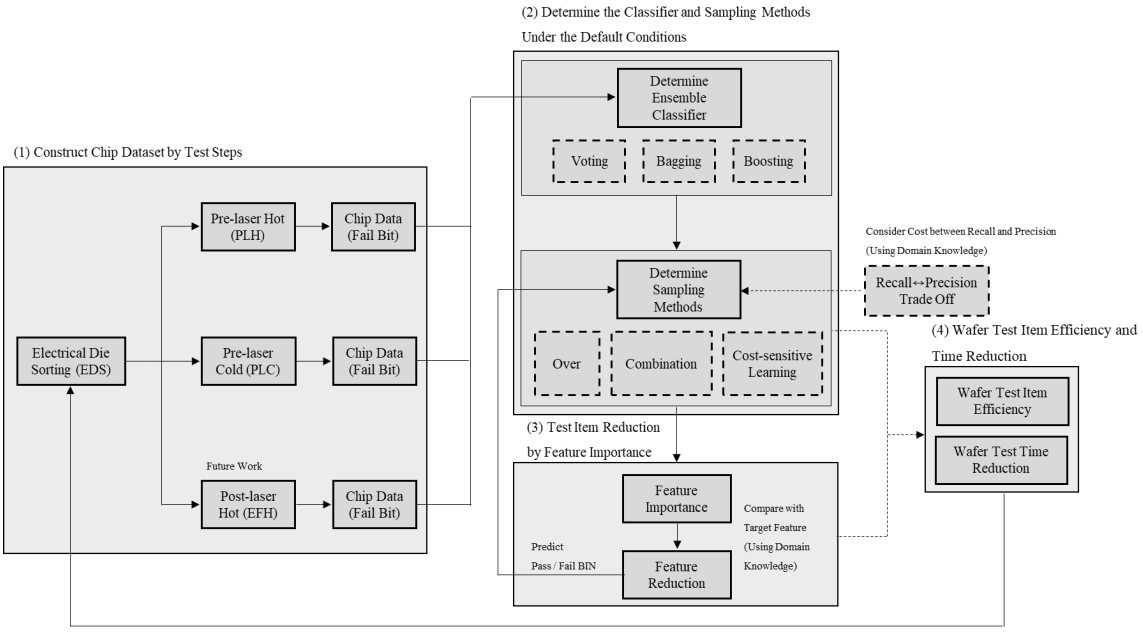


Figure 4. Machine Learning-based Faulty Wafer Classification and Test Item Reduction Framework

같이 크게 (1) 웨이퍼 테스트를 통한 칩 데이터셋 구성, (2) 앙상블 분류기 및 샘플링 방법론 선택, (3) 중요 변수 추출을 통한 분류 및 테스트 항목 효율화, (4) 분류 모델을 이용한 양/불량 칩 예측 및 테스트 시간 단축의 네 단계로 구성된다.

(1) 웨이퍼 테스트를 통한 칩 데이터셋 구성

앞서 언급한 바와 같이 EDS test는 온도에 따라 세 가지 단계로 진행되며 각기 검출하고자 하는 불량의 종류가 다르다. 또한 각 스텝은 50~70여 개의 테스트 아이템으로 구성되어 필드(field)에서 발생 가능한 불량을 선 발현시키기 위해 DRAM 셀들의 성능을 한계점까지 테스트한다. 해당 결점 수는 각 칩 별로 모두 저장되고, 여유 셀의 집합인 리던던시로 리페어하여 대체할 수 없을 경우 해당 칩은 불량 칩으로 BIN 처리되어 이후 테스트는 진행되지 않는다. 단계별 테스트가 완료되면 칩 별 데이터와 웨이퍼 맵은 서버에 저장된다. 칩 별 데이터는 원하는 항목을 임의로 추출 가능하며, 본 연구에서 데이터셋의 변수는 테스트 항목으로만 이루어져 있다.

(2) 앙상블 분류기 및 샘플링 방법론 선택

추출한 데이터셋을 토대로 앙상블 분류기 및 샘플링 방법론을 선택하기 위한 평가를 진행한다. 먼저 default 조건에서 대표적인 5가지 앙상블 알고리즘 (1) voting, (2) bagging, (3) boosting의 분류 성능을 평가하여 최적의 모델을 선택한다. 후속 과업을 진행할 앙상블 분류기를 결정한 후 분류 및 예측 성능을 최대화할 수 있도록 샘플링 방법을 결정하며, 본 실험에서는 크게 세 가지 (1) over-sampling, (2) combination-sampling, (3) 비용민감학습을 사용하였다. 이 과정에서 중요도를 고려하여 새롭게 정의한 평가 지표를 이용하여 불량 칩에 대한 예측 성능은 강화하고 정상 칩의 오분류는 줄이고자 하였다.

(3) 중요 변수 추출을 통한 분류 및 테스트 항목 효율화

앙상블 분류 모델과 샘플링 방법론을 결정하여 분류 과업을 진행 후 알고리즘의 중요 변수 선택법(feature importance selection)을 활용하여 모델 기반 테스트 항목의 중요도를 정의한다. 기본적으로 변수 중요도는 각 모델에 적합한 산출 방식을 활용하였으며, 본 연구에서는 3.1.2절을 통해 선택된 LightGBM 모델의 split-based 변수 중요도를 사용한다. 이 과정에서 제품 엔지니어의 도메인 지식을 활용하여 실제 효율화 하고자 하는 테스트 항목을 비교하여 결정하기 때문에 현업에서 충분히 활용 가능하다.

(4) 분류 모델을 이용한 양/불량 칩 예측 및 테스트 시간 단축

3.1.3절에서 추출한 중요도가 낮은 테스트 항목을 선정 후 해당 변수를 제거(feature reduction)하여 다시 분류 작업을 거쳐 칩들의 양/불량을 예측한다. 이를 통해 실제 효율화 과정에서 소모되는 10일 전후의 평가 기간을 단 몇 분으로 단축시킬 수 있다. 또한 PLH 데이터셋의 경우 현업에 바로 적용 가능한 수준의 좋은 분류 성능을 보이고 있기 때문에 실제 양산 프로그램의 시간 단축 과업에도 즉각적인 반영이 가능하다. 이에 대해서는 5장의 실험 결과에서 자세히 다룬다.

4. 실험 설계

4.1 반도체 웨이퍼 테스트 칩 데이터셋

본 연구에서는 실제 반도체 웨이퍼 테스트로부터 추출된 칩 별 결점 수 데이터를 활용하여 3장에서 제안한 프레임워크의

실험을 진행하였다. 테스트 항목 효율화와 시간 단축이 거의 완료된 구세대 공정 제품에서 고른 웨이퍼 27매에 대한 PLH, PLC 단계의 칩 결점 수와 class 정보로 양/불 이진 값(0 : 정상, 1 : 불량)을 가진다. EFH 단계의 경우 불량을 리페어 한 후의 테스트 결점 수 데이터로 1개 셀에 대한 Fail 이 바로 불량으로 이어지기 때문에 제외하였다. 또한 PLC 데이터셋의 불량 칩은 추가적인 작업을 통해 대표적인 불량 유형인 single bit과 multi bit으로 구분할 수 있도록 하여 상세 분석이 가능하다 (Sridharan and Liberty, 2012). 데이터셋에 대한 min-max normalization, outlier 및 결측치 제거 등과 같은 전처리 과정은 특별히 거칠 필요가 없다. 이는 결점 수 크기 자체도 칩의 양/불에 있어 하나의 구분자이며, 정상적인 EDS test가 진행되었다면 결측치가 존재하지 않고 outlier의 경우 불량 칩으로 구분되기 때문이다. 본 연구에서는 두 가지 단계의 데이터셋을 활용하였으며, 각각에 대한 변수와 관측치는 아래 <Table 1>과 같이 구성되어 있다.

4.2 머신 러닝 이진 분류 모델

본 연구에서는 반도체 칩의 양/불을 분류하고 변수 중요도 선택법을 활용하기 위해 로지스틱 회귀와 더불어 트리 기반의 대표적인 앙상블 이진 분류 알고리즘을 사용하였다. 독립변수의 선형 결합을 통해 사건의 발생 가능성을 예측하고 특정 범주에 속할 확률 값을 계산하는 로지스틱 회귀(James *et al.*, 2013), 다수의 결정 트리(decision tree)를 학습하는 앙상블 방법인 Random Forest(RF)(Ali *et al.*, 2012), 여러 개의 약한 분류기(weak classifier)를 조합하여 가중치 조정을 통해 하나의 강한 분류기(strong classifier)를 합성하는 방법의 Adaptive Boosting(AdaBoost)(Freund and Schapire, 1997), boosting의 개념을 경사 하강 알고리즘을 이용해 최적화하는 Gradient Boosting Machine(GBM)(Friedman, 2001), 그리고 GBM의 확장형태인 LightGBM과 XGBoost가 그것이다.

4.3 불균형 데이터 및 평가 지표

<Table 1>에서 확인할 수 있듯이 반도체 테스트는 정상 칩이 전체 데이터셋의 95% 이상을 차지하는 심각한 데이터 불균형을 지니며, 따라서 소수 범주인 불량 칩에 대한 알고리즘의 예측 성능이 감소하는 문제가 발생한다(TAN *et al.*, 2006). 이와 같은 불균형 문제를 해결하기 위해 유사 연구에서는 Synthetic Minority Over-sampling Technique(SMOTE)를 활용하여 데이터 불균형을 해결하고자 하였다(Park and Kim, 2015). 반면 본 연구에서는 training data에 대해 SMOTE 뿐만 아니라 다양한 over-sampling, combination-sampling, 그리고 비용민감학습 기법(Elkan, 2001)을 모두 적용하여 성능을 향상시키고자 하였으며, 새롭게 정의한 cost-adjusted 평가 지표에 적합하도록 세팅하였다.

비대칭 데이터셋의 평가 지표는 소수 범주에 대한 분류를 간과하지 않도록 combined measure를 사용하여 여러가지 장치를 두는데, 대표적으로 F-Measure, Balanced Accuracy, G-Means 등이 있으며(Bekkar *et al.*, 2013), 해당 지표들은 <Table 2>에 나타낸 혼동 행렬을 기반으로 표현된다.

본 연구에서는 FN과 FP의 중요도를 책정하기 위하여 사원급부터 책임급까지 현업에서 중추 역할을 하는 제품 엔지니어 11명에게 설문조사를 진행하였고 1:3의 중요도 결과를 얻었다. 이는 불량 칩 1개를 더 정분류 할 수 있다면 정상 칩을 3개까지 오분류 하더라도 허용함을 의미하며, 수율이 크게 감소하지 않는다면 고품질을 유지하는 것이 더 중요하는 것을 뜻한다. 해당 결과를 토대로 본 논문에서는 FN의 중요도를 고려하여 아래의 cost-adjusted 평가 지표를 새롭게 정의하였다. 식(1)은 민감도(recall)와 정밀도(precision)의 조화평균인 cost-adjusted F1-score를 의미하며, 식(2)는 민감도(sensitivity)와 특이도(specificity)의 산술 평균인 cost-adjusted balanced accuracy

Table 1. EDS Test Chip Fail Bit Dataset for Experiment

Test Step	Statistics	# of Chip Count	# of Pass Chip	# of Fail Chip	# of Test Item
PLH	Training Set	26,694	25,259	1,435	45
	Validation Set	8,898	8,432	466	45
	Testing Set	8,899	8,427	472	45
PLC	Training Set	25,856	25,264	592	58
	Validation Set	8,571	8,428	143	58
	Testing Set	8,571	8426	145	58

Table 2. Confusion Matrix

		Predicted Class	
		Pass	Fail
Actual Class	Pass	True Negative (TN)	False Positive (FP)
	Fail	False Negative (FN)	True Positive (TP)

score를 뜻한다.

$$Cost-adjusted\ F1\ Score = \frac{2TP}{2TP+FP+3FN} \quad (1)$$

$$Cost-adjusted\ Balanced\ Accuracy\ Score = \frac{1}{2}\left(\frac{TP}{TP+3FN}+\frac{TN}{TN+FP}\right) \quad (2)$$

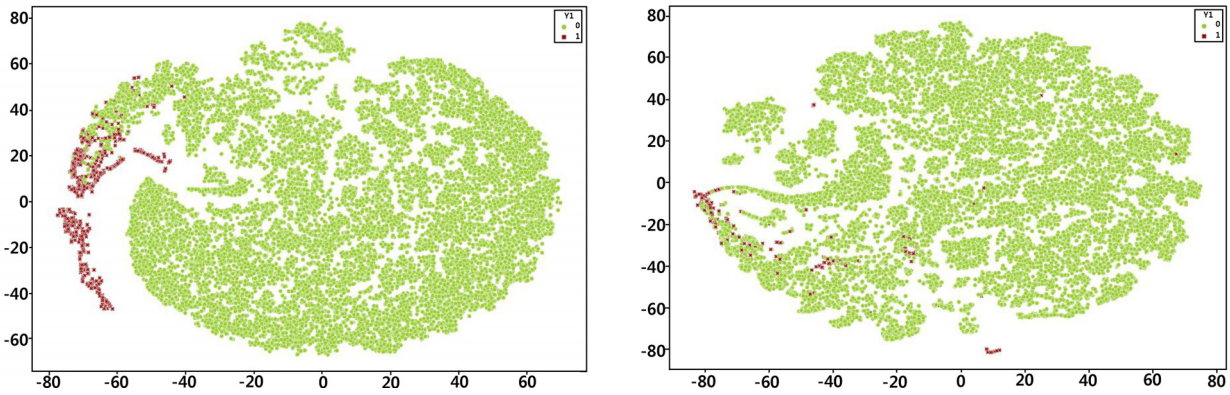
5. 실험 결과

5.1 데이터셋 분포

<Figure 5>는 실험에 사용한 온도에 따른 두 가지 데이터셋을 t-distributed stochastic neighbor embedding(t-SNE)을 통해 차원 축소한 데이터 분포를 나타낸다. PLH 대비 PLC 데이터의 불량 칩 개수가 적고, 정상 칩 사이에 존재하기 때문에 PLC 데이터셋의 이진 분류가 훨씬 더 어려운 과업임을 알 수 있다(정상: 초록색, 불량: 빨간색).

5.2 앙상블 분류기와 샘플링에 따른 분류 성능

두 가지 데이터셋에 3.1절에서 제안한 프레임워크를 수행하였으며, validation data를 이용해 가장 적합한 이진 분류기와 샘플링 방법론을 선택한 후 이후 실험을 진행하였다. <Table 3>에서 볼 수 있듯이 두 가지 cost-adjusted 평가 지표를 기준으로 PLH 데이터셋에서는 모든 분류기가 뛰어난 성능을 보인다. 반면 앞서 언급한바와 같이 PLC 데이터셋에서는 데이터 분포로 인해 성능이 감소하며, 그 중에서 LightGBM이 가장 우수한 성능을 보였다. 본 실험에서는 두 가지 데이터셋 모두 LightGBM을 기본 알고리즘으로 선정하여 이후 실험을 진행하였다. 또한 기존의 평가 지표를 함께 표기함으로써 새롭게 정의한 평가 지표가 중요도를 잘 반영하고 있음을 함께 보였다. 아래 <Figure 6>은 실험에서 선정한 샘플링 방법론으로 소수 범주 데이터를 증가시킨 후 t-SNE를 통해 차원 축소한 분포이다.

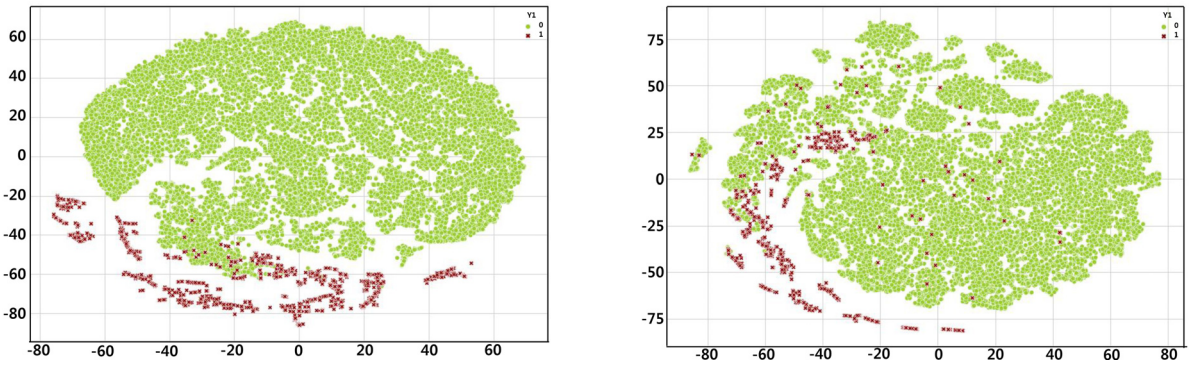


(a) PLH Dataset (b) PLC Dataset

Figure 5. Data Distribution using t-SNE

Table 3. Classification Performance of each Algorithm and Dataset (Validation Set)

Dataset	Algorithm	Hyper-parameters	Cost-adjusted		Original		AUROC
			F1	Balanced Accuracy	F1	Balanced Accuracy	
PLH	Logistic Regression	max_iter=100 penalty='l2'	0.9399	0.9816	0.9495	0.9920	0.9985
	RF	n_estimators=100 criterion='gini'	0.9776	0.9781	0.9924	0.9925	0.9988
	AdaBoost	n_estimators=50 algorithm='SAMME.R'	0.9776	0.9781	0.9924	0.9925	0.9998
	GBM	n_estimators=100 max_depth=3	0.9766	0.9781	0.9914	0.9924	0.9999
	LightGBM	n_estimators=100 boosting_type='gbdt'	0.9776	0.9781	0.9924	0.9925	0.9998
	XGBoost	n_estimators=100 booster='gbtree'	**0.9808**	**0.9812**	0.9935	0.9936	0.9999
PLC	Logistic Regression	max_iter=100 penalty='l2'	0.5112	0.6831	0.7222	0.8171	0.9288
	RF	n_estimators=100 criterion='gini'	0.5910	0.7140	0.8016	0.8459	0.9844
	AdaBoost	n_estimators=50 algorithm='SAMME.R'	0.6217	**0.7432**	0.7940	0.8696	0.9801
	GBM	n_estimators=100 max_depth=3	0.6113	0.7303	0.8016	0.8595	0.9765
	LightGBM	n_estimators=100 boosting_type='gbdt'	**0.6284**	0.7349	0.8221	0.8633	0.9840
	XGBoost	n_estimators=100 booster='gbtree'	0.6066	0.7222	0.8112	0.8529	0.9847



(a) PLH Dataset (b) PLC Dataset

Figure 6. Data Distribution using t-SNE (Sampling, Training Set)

<Table 4>는 PLH 데이터셋에서 LightGBM을 기본 알고리즘으로 각 샘플링 방법론에서 가장 좋은 성능을 보였을 때의 조건과 그 결과를 나타낸다. 하이퍼 파라미터에서 strategy는 다수 클래스 대비 소수 클래스의 샘플링 비율을, neighbors는 가장 가까운 이웃의 개수를, weight는 비용 민감 학습의 cost를 의미한다.

반면 PLC 데이터셋의 경우 샘플링과 비용민감학습을 함께 활용하여 <Table 5>와 같이 F1-score 기준 12.9%, balanced accuracy score 기준 16.0% 성능 향상을 보였다. 해당 결과를 바탕으로 본 연구에서는 좋은 분류 성능을 보이고, 데이터 변형에 대한 위험을 최소화할 수 있는(샘플링 비율이 낮은) 방법론으로 SMOTEENN과 비용 민감 학습을 활용하였다.

Table 4. Classification Performance of LightGBM using Sampling Methods (PLH, Validation Set)

Sampling Methods	Hyper-parameters			Cost-adjusted		Original		AUROC
	Strategy	Neighbors	Weight	F1	Balanced Accuracy	F1	Balanced Accuracy	
Default	-	-	-	0.9776	0.9781	0.9924	0.9925	0.9998
ADASYN	0.14	8.00	1.80	0.9840	0.9842	0.9946	0.9946	0.9998
SMOTE	0.30	6.00	1.00	0.9840	0.9842	0.9946	0.9946	0.9999
Borderline-SMOTE	0.12	2.00	2.00	0.9808	0.9812	0.9935	0.9936	0.9999
SVM-SMOTE	0.18	3.00	1.60	0.9819	0.9841	0.9925	0.9945	0.9999
SMOTE-ENN	**0.12**	**3.00**	**1.20**	**0.9830**	**0.9871**	0.9914	0.9955	0.9998
SMOTE-Tomek	0.12	2.00	1.00	0.9808	0.9812	0.9935	0.9936	0.9998
KMeans-SMOTE	0.30	2.00	2.00	0.9808	0.9812	0.9935	0.9936	0.9999
LORAS	0.10	2.00	1.00	0.9776	0.9781	0.9924	0.9925	0.9998

Table 5. Classification Performance of LightGBM using Sampling Methods (PLC, Validation Set)

Sampling Methods	Hyper parameters			Cost-adjusted		Original		AUROC
	Strategy	Neighbors	Weight	F1	Balanced Accuracy	F1	Balanced Accuracy	
Default	-	-	-	0.6284	0.7349	0.8221	0.8633	0.9840
ADASYN	0.30	4.00	1.40	0.7006	0.8396	0.7848	0.9307	0.9831
SMOTE	**0.18**	**2.00**	**1.60**	**0.7097 (12.9%↑)**	**0.8216 (11.8%↑)**	0.8148	0.9211	0.9850
Borderline-SMOTE	0.60	6.00	1.20	0.7029	0.8335	0.7935	0.9275	0.9861
SVM-SMOTE	0.40	2.00	1.40	0.7014	0.8213	0.8040	0.9209	0.9830
SMOTE-ENN	**0.08**	**3.00**	**1.40**	0.6923 **(10.2%↑)**	**0.8523 (16.0%↑)**	0.7636	0.9369	0.9826
SMOTE-Tomek	0.04	2.00	1.60	0.6869	0.7776	0.8401	0.8943	0.9829
KMeans-SMOTE	0.18	5.00	2.00	0.6748	0.7627	0.8462	0.8842	0.9846
LORAS	0.08	2.00	1.60	0.6505	0.7484	0.8327	0.8737	0.9845

Table 6. Classification Performance (Test Set)

Dataset	Algorithm	Sampling Methods	Cost-adjusted		Original		AUROC
			F1	Balanced Accuracy	F1	Balanced Accuracy	
PLH	LightGBM	-	0.9552	0.9579	0.9839	0.9851	0.9997
		STMOE-ENN Cost-sensitive Learning	0.9645	0.9693	0.9851	0.9892	0.9997
PLC	LightGBM	-	0.6036	0.7205	0.8095	0.8514	0.9766
		STMOE-ENN Cost-sensitive Learning	0.6813 (12.9% ↑)	0.8284 (15.0% ↑)	0.7702	0.9244	0.9777

Table 7. Confusion Matrix of the Best Model (PLH, PLC)

(a) PLH Dataset

PLH Dataset		Predicted Class	
		Pass	Fail
Actual Class	Pass	8,423	4
	Fail	10	462

(b) PLC Dataset

PLC Dataset		Predicted Class	
		Pass	Fail
Actual Class	Pass	8,373	53
	Fail	21	124

앞서 결정한 최적의 이진 분류기 알고리즘과 샘플링 방법론을 이용한 test set 성능은 <Table 6>과 같다. PLH의 경우 default 조건에서 이미 뛰어난 성능을 보이기에 샘플링에 의한 성능 향상이 크지 않았으나 현업 적용 가능성이 충분함을 확인하였다. PLC 데이터셋 또한 12.9% 수준의 성능 향상을 보이며, 극단적 불균형 데이터의 분류 한계를 넘어설 수 있는 가능성을 보여주었다. 또한 <Table 7>의 혼동행렬을 통해 4.3절의 설문조사 결과(불량 칩 1개를 더 정분류 할 수 있다면 정상 칩을 3개까지 오분류 하더라도 허용함을 의미)를 잘 만족함을 확인하였다.

5.3 테스트 아이템 효율화와 시간 단축

본 논문의 궁극적인 목적인 테스트 항목 효율화를 위해 LightGBM에 최적화된 split-based 변수 중요도 선택법을 활용하여 <Table 8>과 같이 중요 변수를 추출하였으며, <Figure 7>과 <Figure 8>은 각각 PLH와 PLC 데이터셋에서 중요도 상위

Table 8. Feature Importance of the Best Model (PLH, PLC)

Dataset	Rank	Top 10			Bottom 10		
		Feature	Importance	Proportion [%]	Feature	Importance	Proportion [%]
PLH	1	X45	218	7.27	X36	0	0.00
	2	X43	202	6.73	X24	0	0.00
	3	X22	154	5.13	X25	0	0.00
	4	X29	148	4.93	X37	0	0.00
	5	X17	137	4.57	X18	3	0.10
	6	X23	132	4.40	X20	6	0.20
	7	X4	122	4.07	X33	7	0.23
	8	X1	121	4.03	X42	9	0.30
	9	X44	119	3.97	X34	16	0.53
	10	X2	117	3.90	X14	20	0.67
PLC	1	X57	271	9.03	X23	0	0.00
	2	X39	211	7.03	X42	0	0.00
	3	X56	164	5.47	X32	0	0.00
	4	X1	157	5.23	X53	0	0.00
	5	X58	130	4.33	X26	1	0.03
	6	X10	122	4.07	X40	2	0.07
	7	X22	116	3.87	X25	5	0.17
	8	X30	102	3.40	X45	9	0.30
	9	X14	101	3.37	X52	9	0.30
	10	X44	101	3.37	X46	11	0.37

아이템과 하위 아이템을 특정 비율로 제거하면서 분류 성능을 측정한 결과를 나타낸다.

중요도 상위 테스트 아이템에 대한 효율화를 진행 시 성능의 급격한 감소가 확인되지만, 중요도가 낮은 아이템 제거 시 성능이 유지됨을 확인하였다. 이는 중요도가 높은 테스트 특정 아이템이 분류 성능에 많은 영향을 미치며, 모델 기반의 변수 중요도가 실제 분류 성능에 밀접한 관계를 가짐을 의미한다. 또한 품질에 영향을 미치지 않는 기준에서 중요하지 않은 테스트 항목을 제품 엔지니어의 경험과 역량이 아닌 객관적인 수치로써 정할 수 있음을 뜻한다.

또한 <Table 9>는 웨이퍼 테스트 항목 효율화를 통해 단축할 수 있는 평가 시간의 감소폭을 확인한 결과이다. Step은 프레임워크의 단계를 의미하며, item count는 해당 단계에서 사용한 모델 혹은 샘플링 방법론의 개수를 뜻한다. 또한 가장 뛰

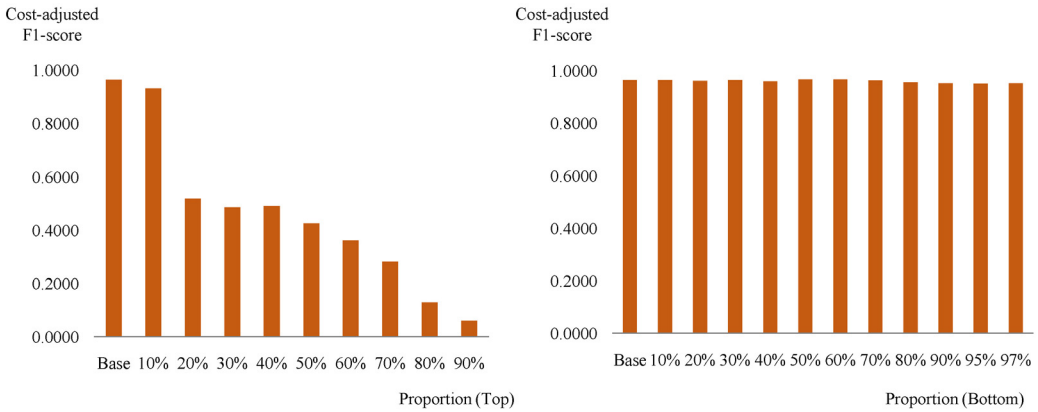


Figure 7. Classification Performance after Feature Reduction (PLH)

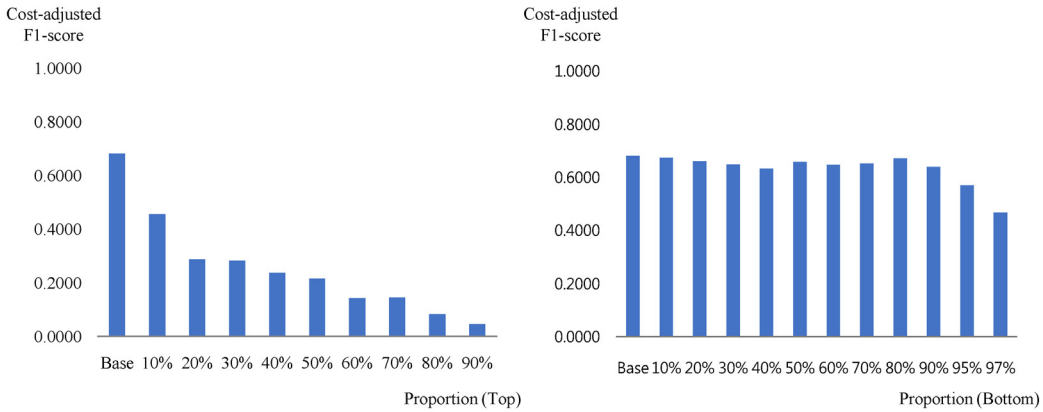


Figure 8. Classification Performance after Feature Reduction (PLC)

Table 9. Time to Perform the Experiments Following the Framework

Dataset	Set	Step	Item Count	Time [s]	Total Time [s]
PLH	Validation	Determine the Classifier	6	15.94	1919.27
	Validation	Select the Sampling Method	5,040	1884.97	
	Test	Classification Performance (Before)	1	0.8	
	Test	Feature Reduction	21	16.76	
	Test	Classification Performance (Final)	1	0.8	
PLC	Validation	Determine the Classifier	6	14.83	3165.21
	Validation	Select the Sampling Method	5,712	3029.46	
	Test	Classification Performance (Before)	1	5.26	
	Test	Feature Reduction	21	110.4	
	Test	Classification Performance (Final)	1	5.26	

어난 이진 분류 성능을 보이는 조합을 결정하기 위해 validation set에 대하여 다양한 분류기와 샘플링 방법의 조합으로 실험을 진행하였으며, 이에 대한 부분은 validation으로 표기하였다. Validation set에 대한 시간은 단 1회만 소모되는 일시적인 시간 소모이며, 알고리즘과 샘플링 방법론 선택 후에는 더 이상 필요하지 않다. 마지막으로 각 단계별 소모되는 시간의 전체 합은 total time으로 표기하였으며, 두 개의 데이터셋에 대해 이진 분류기와 샘플링 조합 선택 시간 포함 시 5084.48초, 제외 시 170.05초라는 짧은 시간에 테스트 효율화에 따른 칩의 양/불을 예측할 수 있었다.

반면 실제 테스트 프로세스의 경우 다음과 같이 크게 (1) 평가 웨이퍼 구성 (2) 효율화 평가 프로그램 셋업 (3) 생산 설비 할당 및 평가 진행 (4) 테스트 아이템 효율화 프로그램 양산 적용 네 단계로 진행된다. Table 1에서 볼 수 있듯이 PLH의 test와 validation data는 17,797개이고, PLC는 17,142개의 칩으로 이루어져 있으며, 이는 웨이퍼 10매에 해당하는 칩 개수이다. 실험에 사용한 데이터셋은 테스트 효율화가 마무리 단계인 제품에서 추출한 것으로, 웨이퍼 1매당 25분 내외의 테스트 시간이 소모된다. 즉, 현업에서 생산 설비를 할당 받아 웨이퍼 10매에 대한 효율화 평가를 진행 시 8시간 이상의 시간이 소모됨을 의미하며, 실제로는 100매 이상의 웨이퍼 평가를 진행하기 때문에 생산성 하락과 병목 현상을 야기한다.

위의 결과를 종합할 때, 본 연구의 궁극적인 목적인 머신 러닝을 기반으로 웨이퍼의 불량을 탐지 및 테스트 항목 효율화가 상당한 수준으로 가능하였음을 보였으며, 이를 토대로 웨이퍼 평가 없이도 객관적으로 평가에 소모되는 시간을 단축하고 생산성을 향상시킬 수 있다는 점에 의미가 있다.

6. 결 론

본 연구에서는 반도체 수요 증가에 따른 생산 리스크를 최소화하기 위한 방안으로 반도체 테스트 항목 효율화와 시간 단축에 대한 연구를 수행하였다. 머신 러닝을 활용하여 높은 불량 웨이퍼 탐지율을 나타내는 최적의 웨이퍼 테스트 아이템 집합을 판별하는 프레임워크를 제시한다. 제안하는 프레임워크는 실제 반도체 웨이퍼 칩 데이터를 바탕으로 성능을 검증하였으며, 두 가지 데이터셋에 대해 각각 의미 있는 결과를 도출하였다. 5장의 PLH 데이터셋의 분류 성능 결과를 바탕으로 머신 러닝이 실제 현업에서 적용 가능한 수준임을 보였으며, PLC 데이터셋의 결과에서는 효과적인 샘플링 방법론을 활용하여 불균형 데이터의 예측 한계를 넘어설 수 있음을 확인하였다. 또한 생산성 하락과 병목 현상을 유발하는 제품 엔지니어의 웨이퍼 테스트 없이도 머신 러닝을 활용하여 짧은 시간에 평가 결과를 예측할 수 있으며, 이를 통해 웨이퍼 테스트 항목 효율화와 시간 단축에 큰 효과가 있음을 보여주었다는데 의의가 있다.

다만, 첫 번째 EDS 테스트로부터 추출되는 PLH 데이터셋으로 후속단의 테스트 없이 칩의 양/불 예측은 적용하기 힘들며, 칩의 결점 수 리페어 후 package test에서의 불량 구분은 불가능했다. 또한 현업에서는 정상과 불량 칩을 유형에 따라 다양하게 나누어 구분하고 있기 때문에 다중 분류의 관점에서는 한계를 가진다. 이를 극복하기 위해 추가적인 image 데이터인 결점 수 맵을 함께 활용하여 다양한 제품, 공정 세대로의 연구 발전이 필요하다.

참고문헌

Ali, J., Khan, R., Ahmad, N., and Maqsood, I. (2012), Random forests and decision trees, *International Journal of Computer Science Issues*, **9**(5), 272.

Baek, D. and Han, C. (2003), Application of data mining for improving and predicting yield in wafer fabrication system, *Journal of Intelligence and Information Systems*, **9**(1), 157-177.

Baek, D. and Nam, J. (2002), Semiconductor yield improvement system using the data mining, *Proceedings of the Korean Operations and Management Science Society*, 296-303.

Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013), Evaluation measures for models assessment over imbalanced data sets, *Journal of Information Engineering and Applications*, **3**(10).

Elkan, C. (2001), The foundations of cost-sensitive learning, *International Joint Conference on Artificial Intelligence*, **17**(1), 973-978.

Freund, Y. and Schapire, R. E. (1997), A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, **55**(1), 119-139.

Friedman, J. H. (2001), Greedy function approximation: A gradient boosting machine, *Annals of Statistics*, 1189-1232.

He, H. and Garcia, E. A. (2009), Learning from imbalanced data, *IEEE Transactions on Knowledge and Data Engineering*, **21**(9), 1263-1284.

Hong, C. and Ahn, J. (2019), The method of parallel test efficiency improvement using multi-clock mode, *Journal of the Semiconductor and Display Technology*, **18**(3), 42-46.

Jahromi, A. H. and Taheri, M. (2017), A non-parametric mixture of gaussian naive bayes classifiers based on local independent features, *IEEE Artificial Intelligence and Signal Processing Conference*, 209-212.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An introduction to statistical learning*, Springer, New York.

Jang, S., Jo, M., Cho, S., and Moon, B. (2018), Defect prediction using machine learning algorithm in semiconductor test process, *Journal of the Korean Institute of Electrical and Electronic Material Engineers*, **31**(7), 450-454.

Jin, C. H., Na, H. J., Piao, M., Pok, G., and Ryu, K. H. (2019), A novel DBSCAN-based defect pattern detection and classification framework for wafer bin map, *IEEE Transactions on Semiconductor Manufacturing*, **32**(3), pp. 286-292.

Kahng, H. and Kim, S. (2020), Self-supervised representation learning for wafer bin map defect pattern classification, *IEEE Transactions on Semiconductor Manufacturing*, **34**(1), 74-86.

Kim, B., Jeong, Y. S., Tong, S. H., Chang, I. K., and Jeongyoung, M. K. (2015), A regularized singular value decomposition-based approach

for failure pattern classification on fail bit map in a DRAM wafer, *IEEE Transactions on Semiconductor Manufacturing*, **28**(1), 41-49.

Kim, E. and Seo, C. (2021), A comparative analysis on export competitiveness of semiconductor industry between Korea and related-significant countries, *The Journal of Asian Studies*, **24**(4), pp. 191-210.

Kim, J., Kim, H., Park, J., Mo, K., and Kang, P. (2019), Bin2Vec: A better wafer bin map coloring scheme for comprehensible visualization and effective bad wafer classification, *Applied Sciences*, **9**(3), 597.

Korea Semiconductor Industry Association. (2021), Global semiconductor industry trends, *Silicon Times*, **601**.

Lee, Y. H., Ham, M., Yoo, B., and Lee, J. S. (2009), Daily planning and scheduling system for the EDS process in a semiconductor manufacturing facility, *The International Journal of Advanced Manufacturing Technology*, **41**(5), 568-579.

Lee, Y. H., Lee, B. K., and Jeong, B. (2000), Multi-objective production scheduling of probe process in semiconductor manufacturing, *Production Planning and Control*, **11**(7), 660-669.

Park, J. and Kim. S. (2015), Predicting package chip quality through fail bit count data from the probe test, *Journal of the Korean Institute of Industrial Engineers*, **41**(4), 408-413.

Smith, R. T., Chlipala, J. D., Bindels, J. F., Nelson, R. G., Fischer, F. H., and Mantz, T. F. (1981), Laser programmable redundancy and yield improvement in a 64K DRAM, *IEEE Journal of Solid-State Circuits*, **16**(5), 506-514.

Sridharan, V. and Liberty, D. (2012), A study of DRAM failures in the field, *IEEE Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, 1-11.

Suh, J. H. and Lee, C. G. (2002), Redundancy analysis simulation for EDS process, *Journal of the Korea Society for Simulation*, **11**(3), 49-58.

Suh, Y. J. and Chung, J. W. (2019), Optimal throughput rate under the price decline of inventory in the semiconductor industry, *Korean Management Science Review*, **36**(1), 67-82.

Tan, P., Steinbach, M., and Kumar, V. (2006), *Introduction to Data Mining*, 1, Addison Wesley.

저자소개

김호영 : 경북대학교 전자공학과에서 2014년 학사학위를 취득하였다. 현재는 삼성전자 메모리사업부에서 책임으로 근무중에 있으며, 고려대학교 산업경영공학부 석사과정으로 재학중이다. 연구분야는 불균형 정형 데이터를 활용한 데이터마이닝과 분류이다.

강필성 : 서울대학교 산업공학과에서 2003년 학사, 2010년 박사 학위를 취득하였다. 이후 현대카드 과장, 서울과학기술대학교 조교수로 근무하였으며, 현재는 고려대학교 산업경영공학부 정교수로 재직 중이다. 연구 분야는 정형 및 비정형 데이터를 활용한 데이터마이닝 및 기계학습 알고리즘 개발 및 제조/IT/공공분야 응용이다.