The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 24, No. 4, pp.57-64, Aug. 31, 2024. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2024.24.4.57
JIIBC 2024-4-9

스마트 팩토리 반도체 공정 데이터 최적화를 위한 향상된 머신러닝 전처리 방법 연구

Enhanced Machine Learning Preprocessing Techniques for Optimization of Semiconductor Process Data in Smart Factories

최승규*, 이승재*, 남춘성**

Seung-Gyu Choi*, Seung-Jae Lee*, Choon-Sung Nam**

요 약 스마트 팩토리의 도입은 제조업 분야에서 객관적이고 효율적인 라인 관리로의 전환을 가져왔다. 그러나 대부분의 회사가 매초 수집되는 수많은 센서 데이터를 효과적으로 사용하지 못하고 있다. 본 연구에서는 이러한 데이터를 활용해 제품 품질을 예측하고 효율적인 생산 공정의 관리를 목표로 한다. 보안 문제로 구체적인 센서 데이터 확인이 불가하여, “SAMSUNG SDS Brightics AI” 사이트의 반도체 공정 관련 학습용 데이터를 확보하여 연구를 진행한다. 머신러닝 모델에서 데이터의 전처리 과정은 성능을 결정짓는 중요한 요소이다. 따라서, 결측값 제거, 이상치 제거, 스케일링, 특성 제거의 전처리 과정을 통해 최적의 센서 데이터를 확보하였다. 또한, 학습 데이터셋이 불균형 데이터를 이루고 있어 오버샘플링 기법을 통해 동일한 비율을 맞추어 모델 평가 전 데이터를 준비하였다. 머신러닝에서 제공되는 다양한 모델 평가로 구한 SVM(rbf) 모델로 높은 성능(Accuracy : 97.07%, GM : 96.61%)을 확인했다. 또한, 동일한 데이터로 학습 시 “SAMSUNG SDS Brightics AI”에서 구현하였던 MLP 모델보다 더 높은 성능을 보인다. 본 연구는 센서 데이터를 활용한 양품/불량품 예측 외에도 부품 주기, 공정 조건 예측 등 다양한 주제에 적용 가능하다.

Abstract The introduction of Smart Factories has transformed manufacturing towards more objective and efficient line management. However, most companies are not effectively utilizing the vast amount of sensor data collected every second. This study aims to use this data to predict product quality and manage production processes efficiently. Due to security issues, specific sensor data could not be verified, so semiconductor process-related training data from the "SAMSUNG SDS Brightics AI" site was used. Data preprocessing, including removing missing values, outliers, scaling, and feature elimination, was crucial for optimal sensor data. Oversampling was used to balance the imbalanced training dataset. The SVM (rbf) model achieved high performance (Accuracy: 97.07%, GM: 96.61%), surpassing the MLP model implemented by "SAMSUNG SDS Brightics AI". This research can be applied to various topics, such as predicting component lifecycles and process conditions.

Key Words : Machine Learning, Preprocessing Methods, Semiconductor Process Data, Smart Factory

*정회원, 인하대학교 소프트웨어융합공학과
**정회원, 인하대학교 소프트웨어융합공학과(교신저자)
접수일자 2024년 7월 28일, 수정완료 2024년 8월 7일
게재확정일자 2024년 8월 9일

Received: 28 July, 2024 / Revised: 7 August, 2024 /
Accepted: 9 August, 2024
*Corresponding Author: namgun99@inha.ac.kr
Dept. of Software Convergence Engineering, Inha University, Korea

Ⅰ. 서 론

반도체, 디스플레이 등과 같은 대기업 제조업 분야의 핵심 목표는 전체 생산 과정에서 일정하고 높은 품질의 제품을 생산하는 것이다. 현재 대부분의 생산 라인은 설비 오류 발생 시 작업자가 수동으로 확인 후 재가동한다. 하지만, 이는 작업자에 따라 조치방법이 다를 수 있어 일정한 제품 품질이 일정하지 않으며, 잘못된 데이터 해석으로 생산성이 하락할 수 있다.[1] 또한, 제조공정에서 모든 설비 데이터는 센서를 통해 FDC Server에 매초 기록되지만, 서버 유지비용으로 인해 에러 발생 당시 단발성 분석 후 일정 기간이 지나 삭제되는 것이 일반적이다.[2, 3] 만약 이러한 데이터를 수집 및 활용해 머신러닝 모델링 후 프로세스를 구축한다면, 공정 중 제품 품질을 예측할 수 있고, 이는 생산성과 수율 부문에서 큰 효과를 나타낼 것이다.[4] 나아가 불량 인자를 파악하고 근본 원인을 수정하는 시퀀스를 구축한다면 완전한 자동화 공정을 실현할 수 있을 것이다. 따라서 본 연구는 Samsung SDS Brightics AI에서 제공되는 반도체 공정 데이터를 가지고 SDS 측에서 이미 구현한 반도체 공정 최적화 모델보다 성능이 향상된 모델을 구현하고자 한다. 이를 통해, 데이터 전처리 과정에서의 특징되는 차이점을 발견하고, 새로운 방식을 추가해 성능을 개선한다. 구체적으로, 본 연구에서는 Samsung SDS Brightics AI의 방식과 달리 결측값 대치, 데이터 스케일링 알고리즘을 변경하고, 모델 성능을 저하시키는 과정인 결측값 비율에 따른 특성 제거, 기존 특성을 이용한 평균값 파생변수 대체, 상관계수 필터링을 사용하지 않았다. 대신 모델 성능을 향상시키기 위해 이상치 제거, 특성 제거 전처리 과정을 새로 추가하였으며, 그 과정을 상세히 나열하였다. 그림 1의 연구를 통해 Smart Factory 분야의 데이터 전처리 과정에 대한 참고자료로 활용될 수 있기를 기대한다.

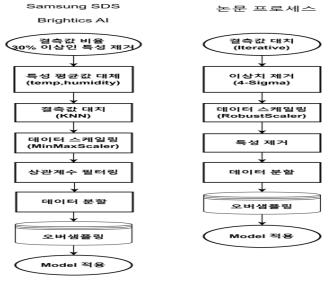

그림 1. 모델 적용 프로세스 비교
Fig. 1. Comparison of Model Application Processes

Ⅱ. 데이터 전처리

1. 데이터 소개

SAMSUNG SDS Brightics AI에서 제공되는 반도체 공정 데이터로 진행한다. 먼저, 학습 데이터는 그림 2와 같이 5개의 공정에 대한 8개의 센서 데이터들로 구성되어 있고, 시간의 흐름에 따라 기록되는 값들로 모두 연속형이라는 특징을 가지고 있다. 이러한 센서 데이터들은 표 1에서와 같다. 크게 공정 온도, 공정 습도, 공정 생산 약품의 유량, 밀도, 점도의 차이. 그리고 공정 반응에서 발생하는 산소, 이산화탄소, 질소의 센서 관측 데이터를 포함해 40개의 특성을 가지고 있으며, 총 16,998행*40열의 데이터를 가진다. 공정 진행이 진행된 생산품의 정상 여부를 확인할 수 있는 타겟 데이터들은 양품 : 0, 불량품 : 1로 정의되고, 13,561:3,437개의 비율을 가진다.[5]

표 1. 데이터 소개
Table 1. Data Introduction

센서명	센서 세부 설명	특성값
Stage#_temp	공정 온도	연속형 (28~32)
Stage#_humidity	공정 습도	연속형 (57~79)
Stage#_flow_deviation	공정 생산력의 흐름 차이	연속형 (-26~26)
Stage#_density_deviation	공정 생산력의 밀도 차이	연속형 (-28~28)
Stage#_viscosity_deviation	공정 생산력의 점도 차이	연속형 (-28~32)
Stage#_o2_deviation	공정 상황에서 발생한 산소 농도 차이	연속형 (-11~11)
Stage#_n_deviation	공정 상황에서 발생한 질소 농도 차이	연속형 (-3~3)
Stage#_co2_deviation	공정 상황에서 발생한 이산화탄소 농도 차이	연속형 (-26~21)
*Target	생산량 양품 여부	0 또는 1

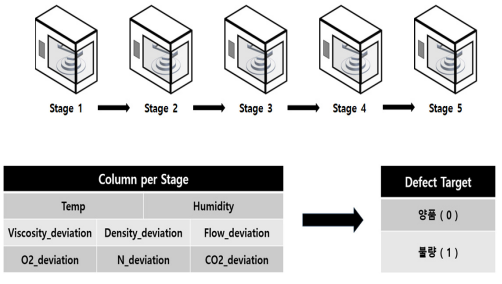

그림 2. 데이터 소개
Fig. 2. Data Introduction

2. 결측값 대치

학습 데이터 내 결측값이 있으면 학습이 불가하다. 따라서 결측값을 대신하는 값을 임의로 채워주어야 한다. 결측값 제거 시 언급한 데이터의 특징 중 수치형(Numeric) 자료에서 강한 성능을 가지는 KNN Imputer와 sklearn.impute에서 제공하는 Iterative Imputer 중 좋은 성능을 보인 Bayesian Estimator를 활용하여 대치 방법을 비교 평가한다.[6, 7]

3. 이상치 제거

실제 공정 데이터는 센서들이 고장 나거나 송/수신 문제로 인해 비정상적인 데이터가 기록되는 경우가 많고, 해당 값들은 올바른 학습 방향에 악영향을 끼칠 수 있다. 그렇기에 센서의 특징과 데이터의 특성을 잘 드러내기 위해 적절한 이상치 제거가 필요할 것이다. 이상치 탐지의 방법을 정하기 위해서는 도메인 지식이 동반되어야 하나, 보안 문제로 인해 공정 도메인은 확인이 불가한 상황이다. 이 점을 감안하여 이상치 탐지 방법을 선정할 것이다. 실제 반도체 FAB 제조공정의 이상을 모니터링할 때 통계적 공정 관리(SPC : Statistical Process Control)를 사용하여 품질 이상 유/무를 판단하고 관리하고 있다. 해당 방법은 데이터들의 평균으로부터 관리 상한(UCL) 및 관리 하한(LCL)을 벗어났을 때 기준으로 이상치 탐지를 하여 공정 불량을 감지한다. UCL과 LCL은 모평균$(\mu) \pm 3\sigma$를 기준으로 지정된다. 따라서 본 논문은 반도체 공정 데이터의 공정 관리 기준인 3σ부터 확률분포가 더 적은 범위를 가지는 4σ와 5σ까지를 기준으로 이상치 제거 방법을 비교 평가한다.[8, 9] 시그마별 분포 확률은 표 2와 같다.

표 2. 시그마별 분포 확률
Table 2. Sigma Distribution Probability

범위	범위 내 확률
$\mu \pm 1\sigma$	68.27 %
$\mu \pm 2\sigma$	95.45 %
$^*\mu \pm 3\sigma$	99.73 %
$^*\mu \pm 4\sigma$	99.9937 %
$^*\mu \pm 5\sigma$	99.999943 %
$\mu \pm 6\sigma$	99.999998 %

4. 데이터 스케일링

모델 학습 간 시간 및 비용을 절감하기 위해 데이터 스케일링 작업은 머신러닝 모델 생성 과정에서 필수적인 단계이다. 데이터 스케일링은 데이터의 각 특징을 일정한 범위 내로 변환하여 모델의 성능을 향상시키고, 학습 속도를 증가시키는 데 중요한 역할을 한다. 본 논문에서는 대표적인 데이터 스케일링 방법 중에서도 특히 많이 사용되는 Standard Scaler, MinMax Scaler, MaxAbs Scaler, Robust Scaler, Normalizer 를 사용한다. 앞서 설명된 전처리 및 모델 조합에서 각 스케일러별 성능을 비교 평가한다.[10]

5. 특성 제거

앞서 [2.1]데이터 소개에서 언급했듯이, 원본 데이터는 16,998*40개의 크기로 구성되어 있으며, 각 열에 다소 많은 결측값이 존재한다. 공정 데이터는 실제 센서 환경에 따라 간헐적으로 비정상적인 출력이 발생하고, 이는 곧 결측값으로 기록된다. 결측값이 많이 기록된다는 것은 공정 진행 중 센서의 성능이 좋지 않다는 의미이며, 이는 데이터의 신뢰성을 저하시킬 수 있다. 신뢰성이 떨어지는 데이터를 사용하면 모델 학습결과가 왜곡되어 성능에 부정적인 영향을 미칠 수 있다.

삼성 측은 표 3과 같이 결측률 30% 이상의 특성들을 제거했으나, 실제로 이러한 조치가 모델 성능을 개선시키는지 명확한 근거가 부족하며, 이는 모델 성능에 악영향을 끼칠 것이라 판단하여 진행하지 않았다.

추가로, 본 논문에서는 특성들의 유기적인 조합에 따른 상호작용이 성능과 크게 연관되어 있기 때문에,[11] 여러 가지 특성의 조합을 테스트할 것이다. 각 공정별 동일한 종류의 센서를 하나의 묶음으로 하여, 총 8개 종류 센서 temp, humidity(humi), co2_deviation (co2_devi), o2_deviation(o2_devi),n_deviation(n_devi),flow_deviation(flow_devi),viscosity_deviation(viscosity_devi), density_deviation(density_devi) 들의 조합을 비교평가 한다. 먼저, 특성 제거를 제외한 전처리 조합을 각 분류 모델에 적용하여, 높은 성능지표를 나타내는 조합을 구한다. 그리고 성능지표 상위 10위에 해당하는 전처리, 모델 조합에 8개 종류의 특성의 전체집합과 공집합을 제외한, 나머지 제거 조합 254개를 비교 평가하여 제거 특성을 도출한다.

6. 데이터 분할

데이터 분할은 모델이 학습된 데이터에 과적합되지 않도록, 검증과 테스트를 통해 일반화 성능을 확인하는 과

표 3. 특성별 결측률
Table 3. Feature Missing Rate

Feature	Stage1		Stage2		Stage3		Stage4		Stage5	
	결측수	결측률	결측수	결측률	결측수	결측률	결측수	결측률	결측수	결측률
temp	0	0	0	0	0	0	0	0	0	0
humidity	0	0	0	0	0	0	0	0	0	0
co2_deviation	*6348	*0.3735	2239	0.1317	2629	0.1547	2629	0.1547	2217	0.1304
o2_deviation	2138	0.1258	1411	0.083	1717	0.101	1717	0.101	3331	0.196
n_deviation	917	0.0539	2352	0.1384	1246	0.0733	1246	0.0733	*5542	*0.326
flow_deviation	*5810	*0.3418	1356	0.0798	2913	0.1714	2913	0.1714	2907	0.171
viscosity_deviation	1407	0.0828	*5622	*0.3307	2904	0.1708	2904	0.1708	*7250	*0.4265
density_deviation	925	0.0544	2506	0.1474	866	0.0509	866	0.0509	2710	0.1594

정이다. 본 논문에서는 sklearn에서 제공하는 “train_test_split” 함수를 활용하여 학습, 검증, 테스트 세트로 데이터 분할을 진행한다. 총 16998개의 데이터를 6:2:2 비율로 분할하여, 학습 데이터 10198개, 검증 데이터 3400개, 테스트 데이터 3400개로 할당하여 평가를 진행한다. 이때 검증 데이터의 양품:불량품의 비율은 2,725:675이고, 테스트 데이터는 2,712:688 개의 불균형한 데이터 비율을 가진다.

7. 오버 샘플링

현재 학습 데이터의 10,198개의 양품/불량품의 비율은 OK:NG=8,124:2,074 개로 현저한 차이가 있어, 모델 성능 향상을 위해 학습 데이터의 불량품의 개수를 늘려주어야 한다. 이 문제를 해결하기 위해 대표적인 오버샘플링 기법인 SMOTE를 사용한다.[12] SMOTE는 불균형 데이터셋에서 소수 클래스 데이터를 생성해 클래스 분포를 균일하게 하여 학습 성능을 향상시키는 기술이다. SMOTE 기법에는 다양한 알고리즘이 있는데, 본 논문에서는 SMOTE, SMOTEN, ADASYN, BorderlineSMOTE, RandomOverSampler, SVMSMOTE, SMOTEENN, SMOTETomek을 활용한다.[13] 표 4에서는 Iterative Imputer를 사용해 결측값 대치를 한 학습 데이터에 오버샘플링 방법들을 적용하여 양품/불량품 비율을 나타내었다. 결측값이 존재하는 상태에서 오버샘플링이 불가하기 때문에 결측값 대치를 진행한 후 오버샘플링 비율을 확인한다.

학습 데이터의 양품/불량품 비율이 한쪽으로 쏠리게 되면, 특정 타겟에 집중적으로 학습되어 분류 성능이 하락할 수 있다. 따라서 표 4에서 볼 수 있듯이 1:1의 비율을 가지지 않는 ADASYN, SMOTENN 알고리즘을 제외한 6개의 오버샘플링 방법을 사용해 비교 평가한다.

표 4. Iterative Imputer 오버샘플링
Table 4. Iterative Imputer Over Sampling

종류	양품	불량품	비율
SMOTE	8121	8124	1 : 1
SMOTEN	8124	8124	1 : 1
RandomOverSampler	8124	8124	1 : 1
ADASYN	8124	8118	1 : 0.99
BorderlineSMOTE	8124	8124	1 : 1
SVMSMOTE	8124	8124	1 : 1
SMOTEENN	5912	7923	1 : 1.34
SMOTETomek	8121	8121	1 : 1

8. 제안된 전처리 방법 적용 모델

위에서 제안된 다양한 데이터 전처리의 조합으로 머신러닝 모델에 적용하기 위해 삼성에서 평가한 분류 모델 중 좋은 성능을 보인 SVM(rbf), MLP Classifier 2개를 사용해 비교 평가한다.[14, 15] 대표적인 성능 지표인 Accuracy(정확도, Acc)와 불균형 데이터 세트에서 평가의 신뢰성을 높일 수 있는 Geometric Mean(민감도, 특이도 기하 평균, GM)을 사용하여 평가한다. 양성 클래스와 음성 클래스를 균형 있게 예측하는지를 확인할 수 있어, 필수로 반영되어야 하는 중요한 지표이다.[16]

위와 같이 데이터 전처리 및 모델을 적용하기 위한 과정은 그림 3과 같다. 그림 3에 나타난 전처리부터 모델 적용까지의 과정별 모든 조합을 평가하였다. 결측값 대치 방법은 KNN, Iterative Imputer 2개를 사용하고, 이상치는 특성 데이터의 3σ와 4σ와 5σ를 기준으로 하는 3개의 방법을 통해 제거하였다. 또한, 대표적인 데이터 스케일링 방법 Standard, MinMax, MaxAbs, Normalization, Robust Scaler 5개의 방법, 특성 제거는 5개 공정에 대한 센서 8개를 종류별로 하나의 묶음으로 하여 평가할 예정이다. 상기 전처리 과정이 모두 이루

어진 데이터를 분할 진행하고, 학습 데이터 10,198개에 대해 오버샘플링을 진행한다. 오버샘플링 시 양품과 불량품이 1:1의 비율을 가지는 6가지 방법을 사용하였다. 최종적으로 분류 모델은 SVM(rbf), MLP Classifier 2개를 사용해 모든 조합을 평가하고 데이터에 적용해 가장 적합한 전처리 방안을 도출하고자 한다.

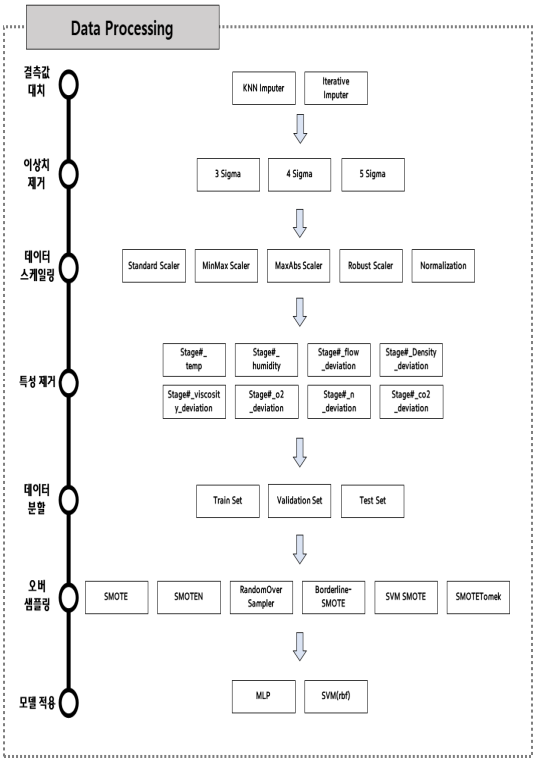

그림 3. 데이터 전처리 및 모델 전체 조합도
Fig. 3. Data Preprocessing and Overall Model Intergration Diagram

Ⅲ. 모델 평가

위 그림 3 에서 특성 제거를 하지 않은 상태에서의 모델 평가를 한 지표는 표 5와 같다. 특성 제거를 제외한 데이터 전처리 및 모델의 조합을 비교 평가한 후 성능 지표를 나타내었다. 특성 제거를 제외한 전처리된 데이터를 기반으로 그림 3과 같이 결측값 대치 2개, 이상치 제거 3개, 데이터 스케일링 5개, 오버샘플링 6개, 분류 모델 2개의 모든 조합 360개의 성능을 평가한 후 상위 10위의 조합을 나타내었다. 불균형 데이터에서 중요한 지표인 GM값을 기준으로 순위를 매겼다.

표 5. 특성 제거 제외한 전처리 및 모델 조합
Table 5. Preprocessing and Model Integration Excluding Feature Elimination

Model	Imputer	remove outlier	scaler	Over sampling	Acc	GM
MLP	iterative	4σ	MaxAbs	svm smote	0.9725	0.9639
MLP	iterative	4σ	MaxAbs	random over sampler	0.9734	0.9593
MLP	iterative	4σ	MaxAbs	smote	0.9728	0.9572
SVM	iterative	4σ	Robust	svm smote	0.9690	0.9554
MLP	iterative	5σ	MaxAbs	svm smote	0.9679	0.9547
MLP	iterative	5σ	MaxAbs	random over sampler	0.9697	0.9546
MLP	iterative	4σ	MaxAbs	smote tomek	0.9728	0.9543
MLP	iterative	5σ	MaxAbs	smote	0.9700	0.9542
MLP	iterative	4σ	MaxAbs	border line smote	0.9675	0.9540
SVM	iterative	4σ	Standard	svm smote	0.9681	0.9537

표 6. 특성 제거 포함한 전처리 및 모델 조합
Table 6. Preprocessing and Model Integration Including Feature Removal

Model	Imputer	remove outlier	scaler	remove feature	Over sampling	Acc	GM
SVM	iterative	4σ	Robust	temp humi o2_devi	svm smote	0.9707	0.9661
MLP	iterative	4σ	Robust	temp humi o2_devi n_devi	svm smote	0.9725	0.96559
MLP	iterative	4σ	standard	temp humi o2_devi	svm smote	0.9760	0.96550
MLP	iterative	4σ	Max Abs	temp humi o2_devi n_devi	svm smote	0.9728	0.9652
MLP	iterative	4σ	standard	temp humi o2_devi n_devi	smote tomek	0.9728	0.9652
SVM	iterative	4σ	standard	temp humi o2_devi	svm smote	0.9754	0.9651
SVM	iterative	4σ	standard	temp humi	svm smote	0.9701	0.9646
MLP	iterative	4σ	Robust	temp humi n_devi	svm smote	0.9754	0.9645
MLP	iterative	5σ	Max Abs	temp humi o2_devi	svm smote	0.9667	0.96449
SVM	iterative	4σ	Max Abs	temp humi o2_devi	smote	0.9743	0.96441

표 5는 특성 제거를 제외한 전처리 및 모델의 조합 결과 중 상위 10위의 성능지표를 나타낸다. 이 중 가장 좋은 조합의 성능은 Acc : 0.9725, GM : 0.9639 으로 나타나는데, 이때의 조합은 Iterative Imputer, 4σ 이상치 제거, MaxAbs Sclaer, SVM SMOTE 오버샘플링을 사용한다. 표 5에서 상위 10위의 성능을 가지는 조합에서 Imputer는 iterative Imputer, 이상치 제거는 4σ와 5σ, 스케일러는 MaxAbs, Robust, Standard를 사용하고, 오버샘플링은 SMOTEN을 제외한 5가지 방법이 차지하는 것을 알 수 있다. 따라서 위와 같이 언급된, 표 5에서 나타난 전처리 방법만을 포함해 특성 제거 조합 254개의 모든 조합 15,240개의 성능평가를 진행하여, 상위 10위의 성능을 표 6에 나타내었다.

표 6에서 특성 제거를 포함하여 전처리 및 모델의 조합을 평가했을 때, 가장 좋은 조합의 성능은 Acc : 0.9707, GM : 0.9661으로 나타난다. 특성 제거 미진행 성능과 동일한 전처리 조건 및 모델 조건에서 비교 시 GM값이 약 1% 상승했고, 특성 제거를 진행함으로써 모델 성능이 개선되었음을 알 수 있다.

표 7. 특성 미제거와 제거 시 최고 성능과 삼성측 성능 비교
Table 7. Comparison of Best Performance with and without Feature Removal to Samsung Performance

Model	Imputer	remove outlier	scaler	remove feature	Over sampling	Acc	GM
SVM	iterative	4σ	Robust	temp humi o2_devi	svm smote	0.9707	0.9661
MLP	iterative	4σ	Max Abs	-	svm smote	0.9725	0.9639
MLP (Samsung)	KNN	-	Min Max	missingvalue 30% more	smote	0.9550	0.9187

표 7은 특성 제거 제외한 경우와 포함한 경우의 최고 성능 조합 및 결과를 나타내며, 최하단 행에 삼성에서 진행한 분류 모델의 성능 결과를 포함하였다. 이를 통해 논문 모델 성능과 삼성 모델의 성능을 비교 평가한다.

표 7의 결과를 보면, 본 논문에서 진행한 전처리 및 모델 조합의 결과가 삼성보다 더 높은 성능을 보인다는 것을 알 수 있다. 본 논문의 최적 조합은 SVM(rbf), Iterative Imputer, 4σ 이상치 제거, Robust Scaler, temp+humidty+o2_deviation 특성 제거, SVM SMOTE를 사용했을 때 Accuracy : 97.07%, GM : 96.61%의 성능을 보인다. 반면 삼성의 최적 조합은 MLP, KNN Imputer, MinMax Scaler, 결측치 30%이상 특성 제거, SMOTE를 사용했으며, 성능은 Accuracy : 95.50%, GM : 91.87%을 보인다. 논문과 삼성 결과 비교 시, 불균형 데이터셋에서 가장 중요한 지표인 GM값이 4.7% 상승해 삼성에서 만든 모델보다 더 뛰어난 성능이 나타나는 결과를 보인다. 추가로, 표 7의 각각의 분류 모델을 활용하여, 테스트 세트 3400개에서 양/불량품의 정상 예측과 비정상 예측 개수를 혼돈행렬(Confusion Matrix)로 시각화했고, 그림 4, 그림 5, 그림 6으로 나타내었다.

그림 4, 그림 5는 각각 특성 미제거와 제거했을 때 최적 모델을 나타냈고, 그림 6은 삼성 측에서 만든 최적 모델의 혼돈행렬 결과이다. 표 7을 통해 나온 동일한 성능 순위임을 알 수 있다. 불량품이 양품의 개수보다 더 적기 때문에, 불균형 데이터에서 불량품 예측 오류를 나타내는 FN(False Negatives)의 개수가 양품 예측 오류를 나

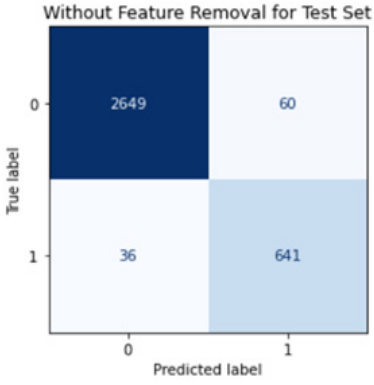

그림 4. 특성 미제거 최적 모델의 테스트셋 혼돈행렬
Fig. 4. Best Model TestSet Confusion Matrix without Feature Removal

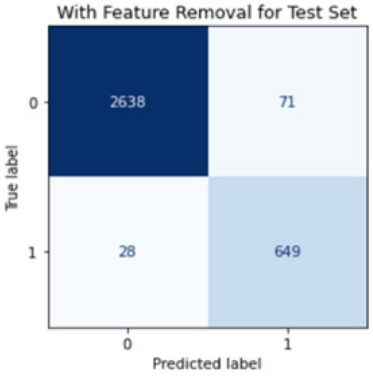

그림 5. 특성 제거 최적 모델의 테스트셋 혼돈행렬
Fig. 5. Best Model TestSet Confusion Matrix with Feature Removal

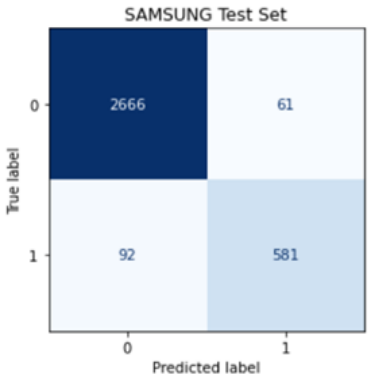

그림 6. 삼성 최적 모델 테스트셋 혼돈행렬
Fig. 6. Samsung Best Model TestSet Confusion Matrix

타내는 FP(False Positives)보다 GM값에 큰 영향을 끼친다. 이로 미루어보았을 때, 삼성 측 모델의 FN(False Negatives)가 92개인 것과 대비하여, 본 논문에서 진행한 특성 미제거와 제거 시의 모델은 FN이 각각 36개와 28개로 확연하게 개선된 것을 알 수 있다.

Ⅳ. 결 론

본 연구와 "SAMSUNG SDS Brightics AI"의 데이터 전처리 방법의 상당한 차이를 확인할 수 있다. 특히, 이 연구에서 가장 좋은 성능을 가지는 전처리 및 모델 조합은 "SAMSUNG SDS Brightics AI"와 달리 모델 성능에 영향을 주는 4σ 기준 이상치 제거를 통해 센서 오계측 값으로 추정되는 값을 제거하였다. 또한, 30% 이상의 결측률을 가진 데이터를 제거하지 않았고, 기존 특성(temp, humidity)를 이용한 평균값 파생변수 대체, 상관계수 필터링을 사용하지 않았다. 대신, 모델 성능 저하에 영향을 주는 특성을 선별하여 제거한 후 모델에 적용하였다. 또한, 데이터 스케일링은 MinMax Sclaer 대신 Robust Scaler를 사용하였고, 오버 샘플링은 SMOTE에서 SVM SMOTE를 사용하며 세부적인 알고리즘의 차이를 보인다. 연구에 사용된 데이터셋은 SAMSUNG SDS Brightics AI" 플랫폼에 등록된 예측모델 MLP Classfier와 비교하여, 학습 데이터를 효과적으로 반영하기 위해 다양한 방법을 통해 데이터 전처리 방법을 변형하여 적용하였다. 이러한 다양한 전처리 방법들을 평가한 결과, 삼성의 모델보다 성능이 개선된 모델을 구현할 수 있었다. 또한, 이렇게 전처리된 데이터를 "SAMSUNG SDS Brightics AI"의 동일 분류 모델인 MLP Classfier에 학습시켜 본 결과, 더 높은 성능을 확인할 수 있었다. 이는 연구에서 채택된 전처리 접근 방식이 더 우수함을 입증하였다.

References

[1] Kwang-Deok Ko, Hyuk-Dae Kwon, Young-Hwan Kwak. "The Impact of Innovative Human Resource Development Activities on Organizational Performance: A Case Study of H Corporation's Semiconductor Backend Process Model", Journal of the Korean Production and Operations Management Society, Vol. 24, No. 4, pp. 491-511, 2013.
DOI: http://doi.org/10.21131/kopoms.24.4.201312.491.

[2] TSUDA, Tomio, et al, "Advanced semiconductor manufacturing using big data", IEEE Transactions on Semiconductor Manufacturing, Vol. 28, No. 3, pp. 229-235, 2015.
DOI: http://doi.org/10.1109/TSM.2015.2445320.

[3] Dae-Geun Ha, Jun-Mo Koo, Dam-Dae Park, & Chong-Hun Han, "APC Technique and Fault Detection and Classification System in Semiconductor Manufacturing Process", Journal of Institute of Control, Robotics and Systems, Vol. 21, No. 9, pp 875-880, 2015.
DOI: http://doi.org/10.5302/J.ICROS.2015.15.0095.

[4] Jae-Wan Yang, Young-Doo Lee, In-Soo Koo, "Sensor Fault Detection Scheme based on Deep Learning and Support Vector Machine", The Journal of the Institute of Internet(JIIBC), Vol. 18, No. 2, pp. 185-195, 2018.
DOI: http://doi.org/10.7236/JIIBC.2018.18.2.185.

[5] Young-seok Seo, "Data Acquired from Various Sensors Installed in Smart Factories Handling Chemical Processes [Data set]", SAMSUNG SDS Brightics AI, 2024.
https://www.brightics.ai/community/knowledge-sharing/data-sets/detail/7098.

[6] Jadhav, Anil, Dhanya Pramod, Krishnan Ramanathan, "Comparison of performance of data imputation methods for numeric dataset", Applied Artificial Intelligence, Vol. 33, No. 10, pp. 913-33, 2019.
DOI: https://doi.org/10.1080/08839514.2019.1637138.

[7] "Imputing missing values with variants of IterativeImputer", Scikit Learn.
https://scikit-learn.org/stable/auto_examples/impute/plot_iterative_imputer_variants_comparison.

[8] Kubo, Tomoaki, Tomomi Ino, Kazuhiro Minami, Masateru Minami, and Tetsuya Homma. "A Statistical Process Control Method for Semiconductor Manufacturing", SICE Journal of Control, Measurement, and System Integration, Vol. 2, No. 4, pp. 246-54, 2009.
DOI: https://doi.org/10.9746/jcmsi.2.246.

[9] CHOI, T. J, et al, "Development of Engineer Change Point Management System in Semiconductor Manufacturing", In Proceedings of the Korean Society of Precision Engineering Conference. Korean Society for Precision

Engineering, pp. 659-660, 2013.

[10] AHSAN, Md Manjurul, et al. "Effect of data scaling methods on machine learning algorithms and model performance", Technologies, Vol. 9, No. 3, Art. 52, 2021. DOI: https://doi.org/10.3390/technologies9030052.

[11] OH, Sejong. "Feature interaction in terms of prediction performance." Applied Sciences, Vol. 9. No. 23, Art. 5191, 2019.
DOI: https://doi.org/10.3390/app9235191.

[12] Young-In Kim, Seon-Jong Kim, Byoung-Chul Kim, Bum-Joo Shin, "Comparing Techniques for Rare Class Classification Problems Using the Overlapped Class Intervals of Data", The Journal of KIIT, Vol. 8, No. 2, pp. 137-144, 2010.

[13] SHARMA, Harsh, GOSAIN, Anushika, "Oversampling methods to handle the class imbalance problem: A review", In International Conference on Soft Computing and its Engineering Applications, Cham, Springer Nature Switzerland, Vol. 1788, pp. 96-110, 2022.
DOI: https://doi.org/10.1007/978-3-031-27609-5_8.

[14] Young-seok Seo, "[Defect Detection] Optimization of Semiconductor Processes in Smart Factories", SAMSUNG SDS Brightics AI, 2024.
https://www.brightics.ai/community/knowledge-sharing/detail/7098.

[15] Hac-Jin Yang, Hyun-Chan Shin, and Seong-Kun Kim, "Prediction of Assistance Force for Opening/Closing of Automobile Door Using Support Vector Machine", Journal of the Korea Academia-Industrial Cooperation Society, Journal of the Korea Academia-Industrial cooperation Society(JKAIS), Vol. 17, No. 5, pp. 364-371, 2016.
DOI: https://doi.org/10.5762/KAIS.2016.17.5.364.

[16] Akosa, Josephine Sarpong, "Predictive Accuracy : A Misleading Performance Measure for Highly Imbalanced Data.", In Proceedings of the SAS global forum, Vol. 12, pp. 1-4, 2017.
https://support.sas.com/resources/papers/proceedings17/0942-2017.pdf.

저 자 소 개

최 승 규(정회원)

- 2017년 ~ : 삼성디스플레이 중소형 사업부 제조기술센터 재직
- 2021년 ~ : 인하대학교 소프트웨어융합공학과 재학
- 관심분야 : 머신러닝, 딥러닝, 컴퓨터 비전, Self-supervised Learning

이 승 재(정회원)

- 2017년 ~ : 진아교역 기술부서 cs팀
- 2017년 ~ : 삼성디스플레이 중소형 사업부 제조기술센터 재직
- 2021년 ~ : 인하대학교 소프트웨어융합공학과 재학
- 관심분야 : 머신러닝, 딥러닝, 이상탐지, 자율주행

남 춘 성(정회원)

- 2005년 : 상명대학교 소프트웨어학과 졸업
- 2011년 : 숭실대학교 컴퓨터학과 석사 졸업
- 2011년 : 성균관대학교 전자전기컴퓨터학과 박사 졸업
- 2011년 ~ 2014년 : 성균관대학교 박사후 연구원
- 2014년 ~ 2016년 : 연세대학교 연구원
- 2016년 ~ 2020년 : 성균관대학교 선임연권 및 책임연구원
- 2020년 ~ : 인하대학교 소프트웨어융합공학과 조교수
- 관심분야 : WSNs, IoT, VANET, HCI, Machin learning, Deep learning

※ 이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역지능화혁신인재양성사업임 (IITP-2024-RS-2023-00259678)