Journal of the Korean Institute of Industrial Engineers
Vol. 51, No. 6, pp. 515-526, December 2025.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2025.51.6.515


# 임곗값 최적화 기반 용접 결함 분류: 클래스 불균형 문제 해소를 중심으로

한민기[1] · 김도희[2] · 배혜림[3†]

[1]부산대학교 산업공학과 / [2]국립창원대학교 인공지능융합공학과
[3]부산대학교 데이터사이언스전문대학원

# Welding Defect Classification Based on Threshold Optimization: A Solution to Class Imbalance Challenges

Min Gi Han[1] · Dohee Kim[2] · Hyerim Bae[3]

[1]Department of Industrial Engineering, Pusan National University
[2]Department of AI Convergence Engineering, Changwon National University
[3]Department of Data Science, Graduate School of Data Science, Pusan National University

Welding defect classification is essential for maintaining the integrity of oil and gas infrastructure, yet it is significantly hindered by severe class imbalance. This study introduces a framework that integrates Random Undersampling and threshold optimization to improve the detection performance of imbalanced datasets. The approach first applies Random Undersampling to reduce majority class samples and rebalance the training set. It then performs post training threshold optimization using multiple evaluation metrics, both with and without a constraint that requires the true positive rate to be greater than or equal to the true negative rate. Across original and resampled datasets, evaluations of various threshold selection strategies, including the default threshold, the class prior threshold, and metric based thresholds, show improved accuracy and a better balance between sensitivity and specificity. The proposed framework increases defect detection while reducing false positives, offering practical guidance for handling other imbalanced binary classification tasks.

***Keywords:*** Welding Defect Classification, Class Imbalance, Random Undersampling, Threshold Optimization

## 1. 서 론

용접은 금속 재료를 열이나 압력을 이용해 접합하는 제조 공정으로 자동, 항공 우주, 조선, 전자 등 핵심 산업 전반에 걸쳐 사용되는 기본 공정이며 생산 설비의 전반적인 품질에 매우 중요한 작업이다(TrIpicchio *et al.*, 2020). 최근 배관 용접 작업에 대한 수요가 증가함에 따라 점점 더 많은 기업과 국가에서도 배관의 품질 확보와 결함 감지 기술이 중요한 과제로 대두되고 있으며, 결함 감지 및 평가 기술은 현재 많이 연구되고 있다(Yang *et al.*, 2021). 실제 용접 결함은 공정에서 복잡한 조건으로 인해 다양한 유형으로 발생하며, 용접 결함의 정확한 감지는 산업 생산의 품질 관리에서 핵심적인 문제이다(Zhang *et al.*, 2019). 그러나 실제 용접 현장에서 이러한 결함은 전체 사례 중 소수에 불과해, 용접 결함 분류를 위한 데이터 세트는 대

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화혁신인재양성사업의 지원(IITP-2025-RS-2020-II201791, 50%)과 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2023-00208999, 50%).
† 연락저자 : 배혜림 교수, 46241 부산광역시 금정구 부산대학로63번길 2 (장전동) 제10공학관 10623호, Tel : 051-510-2733, Fax : 051-512-7603, E-mail : hrbae@pusan.ac.kr
2025년 7월 8일 접수; 2025년 8월 20일; 2025년 9월 2일 수정본 접수; 2025년 9월 2일 게재 확정.

부분 심각한 클래스 불균형 구조를 갖는다.

클래스 불균형을 완화하거나 그 영향을 줄이기 위해서는 데이터를 조정하거나 모델의 학습 및 분류 과정을 개선한다(Leevy *et al.*, 2023). 데이터를 조정하는 방법 가운데 가장 널리 쓰이고 검증된 방법은 1) Random Undersampling(RUS), 2) Random Oversampling (ROS), 3) 합성 소수 클래스 오버샘플링 방법인 Synthetic Minority Over-sampling Technique(SMOTE)가 있다(Leevy *et al.*, 2018). RUS는 다수 클래스 표본을 임의로 제거하여 두 클래스의 비율을 맞추는 방법으로, 불균형을 해소하면서 학습 데이터의 크기를 줄인다(Wongvorachan *et al.*, 2023). 반면에 ROS는 소수 클래스 표본을 그대로 복제해 빈도를 높이는 가장 단순한 오버샘플링 방법이며, 구현이 간단하고 오버샘플링의 기반이 되는 방법이다(Wongvorachan *et al.*, 2023). 이와 비슷하게 SMOTE는 소수 클래스 표본 간 선형 보간을 통해 합성 소수 클래스 표본을 생성함으로써 단순 복제에 따른 과적합을 완화한다(Pradipta *et al.*, 2021).

한편, 모델의 학습 및 분류 과정을 개선하는 방법은 1) 비용 민감 학습(Cost-Sensitive Learning), 2) 임곗값 최적화, 3) 로짓 조정(Logit Adjustment)과 같은 방법이 널리 쓰인다. 비용 민감 학습은 소수 클래스 표본을 잘못 분류했을 때 더 큰 페널티를 주고, 다수 클래스 오류에는 상대적으로 작은 페널티를 부과함으로써 불균형을 완화하는 방법이다(Araf *et al.*, 2024). 또한, 임곗값 최적화는 모델이 분류한 확률을 기반으로 특정 성능 지표가 최대가 되도록 임곗값을 최적화하여 불균형을 완화하는 방법이다(Hancock *et al.*, 2022). 마지막으로 로짓 조정은 모델의 선형 출력값(로짓)에 클래스별 사전 확률을 기반으로 계산된 값을 더해, 분류 기준을 소수 클래스 방향으로 이동시키는 방법이다(Wang *et al.*, 2023).

본 연구에서는 데이터 분포 편향을 해소하기 위해 RUS를, 모델의 학습 및 분류 과정을 개선하기 위해 임곗값 최적화를 적용하여 결합하였다. RUS는 데이터 크기를 줄여 소수 클래스를 더 잘 학습할 수 있게 하고 과적합을 줄이며, ROS와 달리 분류 모델 훈련 시간을 단축할 수 있다(Wongvorachan *et al.*, 2023; Hasanin *et al.*, 2019). 추가적으로, ROS 및 SMOTE 계열 기법이 소수 클래스 샘플을 복제하거나 합성하는 과정에서 노이즈와 모호한 경계를 유발하여 분류기의 과적합으로 이어질 수 있고, 이에 비해 RUS는 이러한 위험이 상대적으로 낮아 보다 더 안정적인 일반화 성능을 확보할 수 있다(Joloudari *et al.*, 2023). 또한 임곗값 설정은 분류 성능에 결정적 영향을 미치는데(Johnson and Khoshgoftaar, 2021), 임곗값을 높게 설정하면 실제 결함 표본을 놓치기 쉽고, 낮게 설정하면 정상 표본이 결함으로 오 분류될 가능성이 커진다. 따라서 임곗값 최적화를 적용해 최적 임곗값을 탐색 및 적용함으로써 정상 및 결함 간 오 분류 균형을 맞추고 분류 정확도를 개선한다(Leevy *et al.*, 2023).

이 두 방법의 결합은 데이터 분포의 편향과 분류 기준의 치우침을 동시에 완화한다. 또한, 대부분의 분류 모델은 기본 임곗값으로 0.5를 사용하지만(Johnson and Khoshgoftaar, 2021), 클래스 불균형 상황에서는 이 값이 최적 기준이 아닐 수 있다. RUS에 기반한 클래스 분포 변화는 임곗값 설정에 영향을 준다는 것을 실험적으로 확인하였다(Leevy *et al.*, 2023). 실제 결함 표본 중 모델이 결함으로 올바르게 분류한 비율(TPR)이 실제 정상 표본 중 모델이 정상으로 올바르게 분류한 비율(TNR) 이상으로(TPR ≥ TNR) 유지되도록 제약을 두면 민감도(TPR)와 특이도(TNR) 간 균형을 개선하고, 대부분의 표본이 음성으로 분류되는 것을 방지한다(Hancock *et al.*, 2022). 위 연구 결과를 바탕으로, 본 연구에서는 다음과 같은 가설을 설정하였다.

가설 1: 클래스 불균형 데이터에서 기본 임곗값 0.5가 적합하지 않다.
가설 2: RUS 비율이 증가할수록 최적 임곗값이 상승한다.
가설 3: TPR ≥ TNR 제약은 민감도와 특이도 간 균형을 개선한다.

위 가설의 검증을 통해, 본 연구가 가지는 기여는 다음과 같다. 클래스 불균형을 보이는 배관 용접 결함 분류 문제에 대해 RUS와 임곗값 최적화를 결합한 방법론을 제안한다. RUS로 클래스 분포를 조정한 뒤, 지표별 최적 임곗값과 TPR ≥ TNR 제약 적용 여부를 비교하여 데이터 분포 편향과 분류 기준의 치우침을 종합적으로 평가하였다. 실험 결과, 본 방법론은 분류 성능을 유의미하게 개선하고, TPR과 TNR 간 균형을 달성하였다. 이에 따라 용접 결함 분류를 넘어 다양한 클래스 불균형 이진 문제에도 활용할 수 있는 실용적 지침을 제공한다.

제2장에서는 클래스 불균형 완화와 용접 결함 분류의 선행 연구를 정리하고 기존 연구의 한계를 도출한다. 제3장에서는 본 연구에서 제안하는 데이터 수집 및 통합, 전처리, RUS 기반 클래스 비율 조정, 분류기 학습 및 선정, 임곗값 최적화 알고리즘을 포함한 통합 프레임워크 및 실험 절차를 설명한다. 제4장에서는 실험 결과를 분석하고 앞서 제시한 가설들을 검증한다. 마지막으로 제5장에서는 본 연구의 결론 및 한계점과 향후 연구를 제시한다.

## 2. 관련 연구

2.1절에서는 클래스 불균형 완화 방법, 2.2절에서는 용접 결함 분류에 대한 선행 연구를 정리하였다.

### 2.1 클래스 불균형 완화 방법 연구

최근 클래스 불균형 문제를 해결하기 위해 제안된 주요 방법들을 <Table 1>에 정리하였다. 먼저, 용접 결함 분류 분야에서는 Hou *et al.*(2019)이 Random Undersampling, Random Oversampling, SMOTE를 적용하여 비교 분석하였으며, Park *et al.*(2019)은 이미지 변환 및 왜곡 기반 데이터 증강과 각 배치

**Table 1.** Summary of Class-Imbalance Handling studies

| Author(Year) | Level | Utilized Method | Stage | Domain | Hybrid Strategy |
|---|---|---|---|---|---|
| Hou *et al.*(2019) | Data | RUS, ROS, SMOTE | Pre-processing | Welding | X |
| Park *et al.*(2019) | Data | Data Augmentation + Class-specific batch | Pre-processing | Welding | X |
| Zhang *et al.*(2019) | Data | Data Augmentation | Pre-processing | Welding | X |
| Johnson and Khoshgoftaar(2021) | Algorithm | Threshold Optimization | Post-processing | Finance | X |
| Hancock *et al.*(2022) | Data | Random Undersampling | Pre-processing | Medical | X |
| Kini *et al.*(2022) | Data | Random Oversampling | Pre-processing | Medical | X |
| Arafa *et al.*(2022) | Data | Reduced Noise SMOTE | Pre-processing | Biology, Ecology | X |
| Hancock *et al.*(2022) | Algorithm | Threshold Optimization | Post-processing | Finance, Medical | X |
| Tian *et al.*(2024) | Algorithm | Focal Loss | In-processing | Finance | X |
| Ours | Hybrid (Data +Algorithm) | RUS + Threshold Optimization | Pre/Post-processing | Welding | O |

의 정상과 결함 이미지의 비율을 일정하게 구성하는 방법인 class-specific batch sampling을 통해 클래스 비율을 보정하였다. 또한 Zhang *et al.*(2019)은 노이즈 추가 및 회전 기반 데이터 증강을 적용하여 불균형을 완화하였다.

본 연구의 도메인과는 다른 분야에서 사용된 데이터를 조정하는 방법으로는 Random Undersampling(Hancock *et al.*, 2022), Random Oversampling(Kini *et al.*, 2022), 그리고 RN-SMOTE (Arafa *et al.*, 2022)와 같은 합성 오버샘플링 방법이 있다. 이러한 방법들은 학습 전에 클래스 분포를 조정하지만, Random Undersampling은 다수 클래스의 정보를 손실할 수 있고, Random Oversampling과 SMOTE는 과적합 위험과 데이터 세트가 증가하는 한계점이 존재한다.

모델의 학습 및 분류 과정을 개선하는 방법은 Focal Loss를 손실 함수로 적용하여 이를 동적으로 조정해 소수 클래스에 가중치를 부여하는 접근(Tian *et al.*, 2024)과, Threshold Optimization으로 결정 임곗값을 조정해 클래스 불균형을 완화하는 접근(Johnson and Khoshgoftaar, 2021; Hancock *et al.*, 2022)이 있다. Hancock *et al.*(2022)은 불균형 데이터에서 작은 임곗값 이동이 성능을 크게 좌우한다는 점을 검증했고, Johnson and Khoshgoftaar(2021)은 TPR $\geq$ TNR 제약을 두고 지표별 최적 임곗값을 탐색해 기본값 0.5보다 우수한 성능을 확인했다. 그러나 이와 같은 방법들은 데이터 분포 자체를 조정하지 않아 원본 데이터 분포 차이가 있다는 한계가 있다.

### 2.2 용접 결함 분류 연구

최근 X-ray 이미지와 전류 및 전압 센서 데이터, 공정 데이터 등 다양한 데이터원을 활용한 기계학습 및 딥러닝 기반 용접 결함 분류 연구가 활발히 진행되고 있다. 이미지 데이터에 YOLO 계열 모델을 적용한 Yang *et al.*(2021)과 Wang *et al.*(2023)은 실시간 결함 검출 정확도를 개선하였으나, 정상 및 결함 클래스 간 불균형을 해소하지 못하였다. 센서 데이터 기반 접근법을 제시한 Moinuddin *et al.*(2021)은 전류, 전압 신호에서 통계적 특징을 추출한 뒤 정상 데이터와 결함 데이터의 비율을 일대일로 맞춘 후 이를 기반으로 Decision Tree와 SVM 성능을 비교하였으나, 임곗값 최적화 절차는 포함하지 않았다. Hahn *et al.*(2023)은 Autoencoder-LSTM 구조를 통해 이진 분류를 수행하였으나, 클래스 불균형 해소와 임곗값 최적화는 수행하지 않았다. 공정 변수를 이용한 Liu *et al.*(2024)와 정상 데이터와 결함 데이터 간 균형을 맞춘 Vasan *et al.*(2024) 역시 결정 임곗값을 조정하지 않은 채 용접 결함을 분류하였다. <Table 2>에서 볼 수 있듯이, 기존 연구들에서 용접 결함을 분

**Table 2.** Summary of Welding Defect Classification studies

| Authors(Year) | Utilized Data | Utilized Method | Class Balancing | Threshold Optimization |
|---|---|---|---|---|
| Yang *et al.*(2021) | Image data | YOLOv5 | X | X |
| Moinuddin *et al.*(2021) | Sensor data | Decision Tree, SVM | O | X |
| Wang *et al.*(2023) | Image data | YOLOv5 | X | X |
| Hahn *et al.*(2023) | Sensor data | Autoencoder, LSTM | X | X |
| Liu *et al.*(2024) | Process data | Hybrid Simulated annealing | X | X |
| Vasan *et al.*(2024) | Image data | GLCM, GLDM, MLP, etc. | O | X |
| Ours | Process data | CatBoost | O | O |

류하기 위해 다양한 방법론이 적용되었으나 데이터 수준의 클래스 분포 조정(Class Balancing)과 임곗값 최적화(Threshold Optimization)를 동시에 반영한 연구는 없었다.

따라서 본 연구에서는 클래스 불균형을 보이는 배관 용접 결함 분류 문제를 위해 다음과 같은 차별성을 가진다. 2.1절에서 언급한 과적합 위험, 데이터 세트 증가, 원본 데이터 분포 차이의 한계점을 해결하기 위해 RUS로 클래스 분포를 조정한다. 본 절의 한계점인 클래스 분포 조정과 임곗값 최적화를 동시에 반영한 연구가 없다는 선행 연구를 바탕으로 RUS와 임곗값 최적화를 결합한 방법론을 제시한다. 이를 통해 본 연구는 선행 연구의 한계점을 효과적으로 해결하여, 용접 결함 분류에서의 클래스 불균형 완화 기법에 대한 새로운 접근법을 제시할 수 있다.

## 3. 제안 방법론

본 연구의 프레임워크는 <Figure 1>과 같다. 다양한 출처의 데이터를 수집 및 통합한 뒤, 도메인 지식에 기반한 변수 선택과 전처리를 수행한다. 이어서 학습용 데이터와 평가용 데이터로 분할하고, 학습용 데이터에 대해 특정 비율로 RUS를 적용한다. 이렇게 준비된 학습 데이터로 3개의 분류기에 대해 학습을 진행하고 가장 성능이 좋은 모델을 선택한다. 마지막으로 선택된 모델을 바탕으로 검증 및 평가와 임곗값 최적화를 수행하여 최종적으로 용접 결함을 분류한다.

### 3.1 Data Collection

본 연구에서 사용한 배관 용접 결함 통합 데이터는 세 개의 원본 데이터를 단계적으로 병합하여 완성하였다. 용접 공정 및 검사 이력 데이터는 2021년부터 2023년까지 3개년의 현장 기록 데이터로, 용접방법, 용접자세 등과 같은 작업 난이도를 표현하는 공정 변수뿐만 아니라 용접사 나이, 경력 등 인적 변수를 포함한다. 이를 NDT(비파괴검사) 기록 데이터와 ITP 번호를 고유 키로 사용하여 1차로 병합한 다음 용접사 등급 데이터의 복합 키를 적용해 2차 병합을 수행하였다. 이 두 단계 조인을 통해 공정 변수, 비파괴 검사 결과, 그리고 용접사 개인 지표가 통합된 배관 용접 결함 데이터를 통합하였다.

본 연구의 데이터 세트는 불량이 발생할지, 안 할지 분류하는 대표적인 이진 분류(binary classification) task 문제로 종속변수는 불량 이슈이다. 불량 이슈는 불량 매수가 0보다 크면 1(결함), 그렇지 않으면 0(정상)으로 부여된다. 전체 표본 가운데 불량인 경우는 약 2.79%에 불과하며, 양성 클래스와 음성 클래스 간의 불균형이 심각하다는 특징이 있는 데이터이다.

### 3.2 Feature Engineering

본 연구는 도메인 지식을 바탕으로 재료 및 치수, 용접 공정, 작업자 능력, 검사의 네 가지 관점에서 총 19개의 변수를 최종 선정하였다. Gao *et al.*(2023)은 배관 용접 작업의 경우, 파이프의 두께, 직경, 공차 및 재질 등급과 같은 기하 및 물성 특성이 용접부의 잔류응력 분포와 결함 발생을 결정짓는다고 언급했고, Muthmann *et al.*(2009)은 고강도 강재일수록 용접이 어렵고, 예열 및 층간 온도를 관리해야 하므로 용접 난이도가 더욱 가중된다고 언급했다. 이를 근거로 재질, 두께 등과 같은 재료 및 치수 관련 변수들을 선정했다. Apeh *et al.*(2023)은 전류, 전압, 조인트 형상, 재료 등 공정 매개변수가 용접부 품질과 열영향부 특성을 결정한다고 밝혔고, Cheng *et al.*(2021)은 용접자세가 품질 편차에 명백한 영향을 준다고 하여 이를 바탕으로 유체종류, 용접자세 등 용접 공정 변수들을 선정하였다. 또한

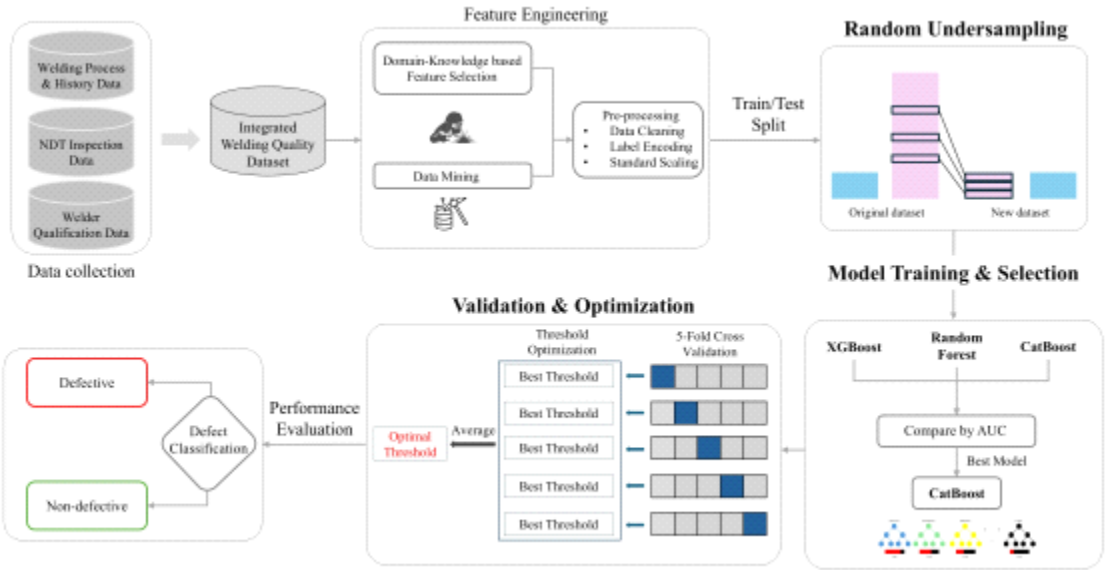



**Figure 1.** Overall Framework of the Proposed Method

Kumar *et al.*(2018)은 숙련된 용접사가 용접 간격을 일정하게 유지해 전압 변동을 최소화하고 균일한 용접 품질을 달성하는 반면, 비교적 덜 숙련된 용접사는 변동이 크다는 사실을 입증했고 해당 연구를 바탕으로 경력 등 숙련도 관련 변수를 선정하였다. 마지막으로 Shaloo *et al.*(2022)의 연구에서는 초음파, 레이저 등 NDT 방법이 실시간 비파괴 평가로 비용 효율적인 결함 탐지를 가능하게 한다는 연구를 근거로 검사 방식과 검사 길이 등을 변수로 포함하였다.

변수 선정 후 전처리 과정에서는 결측치나 의미 없는 값을 제거하여 입력 오류로 인한 왜곡을 방지하였다. 그 후 범주형 변수는 Label Encoding(Hancock *et al.*, 2020)을 통해 고유한 정수로 변환하였다. 또한, 수치형 변수는 평균을 0, 표준편차를 1로 변환하는 표준화(Standardization)를 적용하여 변수 간 스케일 차이가 학습에 미치는 영향을 완화함으로써 학습 안정성을 확보하였다(Ahsan *et al.*, 2021). 검증 데이터의 정보가 학습 과정에 반영되는 것을 방지하기 위해, k-fold cross-validation (k=5)은 각 fold마다 학습용 데이터에 대해 정규화를 수행하였다. 이와 같은 전처리 과정을 통해 잡음과 스케일 편차를 완화했다.

### 3.3 Random Undersampling

본 연구에서는 데이터의 클래스 불균형을 제어하기 위해 imbalanced-learn 라이브러리의 RandomUnderSampler 모듈(Lemaitre *et al.*, 2017)을 활용하였다. 초기 클래스 비율(Original)을 유지하는 조건을 제외하고, 소수 클래스와 다수 클래스 비율을 1:2(RUS = 0.5), 1:3.3(RUS = 0.3), 1:10(RUS = 0.1)의 세 가지 수준으로 설정하여 실험을 수행하였다. 실험마다 Stratified k-fold Cross Validation 방법을 적용하여 전체 데이터의 80%에 대해 학습을 수행하며, k에 5를 할당하여 4개의 폴드(Fold)를 훈련에 사용하고 1개의 폴드(Fold)를 테스트에 사용한다. 계층화(Stratification)를 통해 각 폴드(Fold)별 클래스 비율이 원본 분포를 유지하도록 구성하였다.

단일 RUS 수행은 무작위성으로 인해 성능의 변동성을 유발하므로, 이를 완화하기 위해 반복적으로 수행하였다. 반복 횟수 {5, 10, 20}에 대해 성능 변동성을 비교한 결과(<Figrue 2>), 5회에서 10회로 증가시 변동성이 약 5.0% 감소한 반면, 10회에서 20회로 증가시 추가 감소 폭은 약 4.7%로 감소하였다. 또한

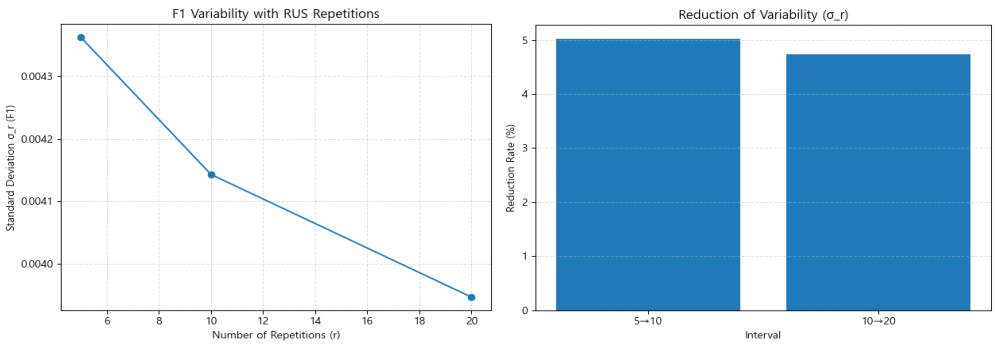
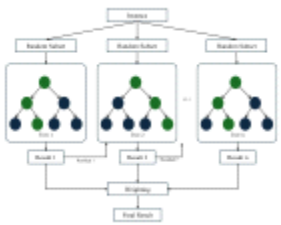
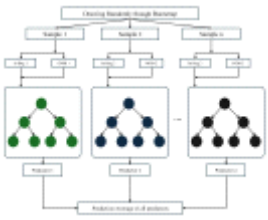



**Figure 2.** Effect of RUS Repetitions on Performance Variability

**Table 3.** Comparison of Classification Algorithms

| | XGBoost | CatBoost | Random Forest |
|---|---|---|---|
| Author (Year) | Chen and Guestrin (2016) | Prokhorenkova *et al.* (2018) | Breiman (2001) |
| Structure |  |  |  |
| | • Extended GBDT<br>• Sparsity-aware split<br>• Weighted quantile sketch<br>• Cache optimization | • Order-based encoding<br>• Ordered Boosting<br>• Symmetric tree structure | • Bagging-based decision-tree ensemble<br>• Bootstrap sampling<br>• Random feature subsets |
| Advantages | Provides fast training and high classification accuracy on large-scale, sparse data | Delivers stable performance on datasets with many categorical features | Reduces variance for overfitting resistance; simple implementation and easy interpretation |

20회 반복은 계산 비용이 크게 증가하는 한계가 있다. 따라서 계산 효율성과 성능 안정성 간 균형점을 고려하여 RUS 반복 횟수를 최종적으로 10회로 설정하였다. 이를 통해 동일한 RUS 비율 및 임곗값 최적화 조건에서 균형 잡힌 성능을 보이는 분류기 및 임곗값 조합을 선정한다.

### 3.4 Classification

본 연구에서는 Xgboost, Random Forest, CatBoost 세 가지 트리 기반 앙상블 모델을 적용하였다. Xgboost는 확장형 그래디언트 부스팅 결정트리 알고리즘으로 대규모 데이터에서도 빠른 학습을 제공한다. 한편 Catboost는 부스팅 알고리즘으로 범주형 변수가 많은 데이터에서 안정적인 성능을 보여준다. 또한 Random Forest는 배깅(bagging)기반 결정트리 앙상블로, 최종 예측을 다수결로 통합함으로써 모델 간 상관을 줄여 분산을 낮추고 강건한 성능을 보여준다. 세 모델의 구조 및 특징은 <Table 3>과 같이 정리된다.

### 3.5 Threshold Optimization

**Algorithm 1:** Threshold Optimization

**Input:**
- $y$ ground truth labels
- $\hat{y}$ classifier output probabilities
- $f$ optimization function, $.e.g.$ f-measure
- $c$ flag controlling whether optimization is constrained

**Output:** $\lambda$ Optimized threshold

**Initialize**
  best_score ← $-\infty$
  best_threshold ←0.5
tprs, fprs, thresholds ← roc_curve($y, \hat{y}$)
tnrs ←1 - fprs
scores ← $f(y, \hat{y})$
**If** $c$ = True **then**
  valid_idx ← tprs ≥ tnrs
  thresholds ← thresholds[valid_idx]
  scores ← scores[valid_idx]
**End If**
**For** i ←1 to length(thresholds) do
  **If** scores[i] > best_score **then**
    best_score ← score[i]
    best_threshold ← thresholds[i]
  **End If**
**End For**
**Return** best_threshold

학습 데이터에서 분류기가 출력한 모든 고유 확률값을 임곗값 후보로 설정하고, 지정된 평가 지표를 최대화하는 최적 임곗값을 탐색하는 알고리즘은 Algorithm1로 표현된다. 구체적으로, 실제 라벨 $y$와 예측 확률 $\hat{y}$을 roc_curve($y, \hat{y}$) 함수에 입력하면, 함수가 각 임곗값에 대한 TPR, FPR, thresholds 배열을 출력한다. 이후 TNR = 1 - FPR을 계산하고, 각 후보 임곗값마다 F-measure, G-mean, MCC, Precision 등 선택한 지푯값을 scores 배열에 저장한다.

알고리즘은 제약 조건 여부를 제어하는 Flag($c$)를 통해 TPR ≥ TNR 제약 조건을 적용할 수 있으며, Flag($c$) = True인 경우 제약 조건을 만족하는 임곗값 후보 중 지푯값이 최대인 임곗값을 최종 선택한다. 이로써 양성 클래스 민감도(TPR)가 특이도(TNR)보다 낮아지는 편향적 임곗값을 방지할 수 있다. 위에서 언급한 4가지 지표에 대해 Flag($c$)를 True/False로 설정하여 총 8개의 최적 임곗값을 도출하고 기본 임곗값 0.5와 원본 데이터의 양성 사전 확률(Class Prior)을 추가하여 총 10개의 임곗값에 대해 비교한다. 다음 절에서는 실험 설계 및 평가 지표와 최적 임곗값에 따른 성능 결과를 제시한다.

## 4. Result and Analysis

### 4.1 실험 설계 및 평가 지표

본 연구에서는 선행 연구에서 많이 사용하는 세 가지 트리 기반 분류기-Xgboost(Chen and Guestrin, 2016), CatBoost (Prokhorenkova *et al.*, 2018), Random Forest(Breiman, 2001)-의 성능을 python 실험 환경에서 비교하였다. 모델 간 성능 비교는 임곗값 설정의 영향을 받지 않은 ROC(Receiver Operating Characteristic) 곡선 아래 면적, 즉 AUC(Area Under the Curve)로 평가하였다. <Table 4>는 실제 클래스(Actual class)와 분류 클래스(Predicted class)의 조합에 따라 True Positive, True Negative, False Positive, False Negative를 정의한다. 이를 바탕으로 True Positive Rate(TPR), True Negative Rate(TNR), False Positive Rate(FPR), False Negative Rate(FNR) 네 가지 비율 지표를 산출하였다(식 (1)-(4)).

**Table 4.** Definition of True Positive, True Negative, False Positive, And False Negative

| | | Predicted class | |
|---|---|---|---|
| | | Defective | Good |
| Actual class | Defective | True positive (TP) | False negative (FN) |
| | Good | False positive (FP) | True negative (TN) |

$$TPR = \frac{TP}{TP+FN} \tag{1}$$

$$TNR = \frac{TN}{TN+FP} \tag{2}$$

$$FPR = \frac{FP}{TN+FP} \tag{3}$$

$$FNR = \frac{FN}{TP+FN} \tag{4}$$

3.5절에서 언급했던 임곗값 최적화의 기준으로 사용된 성능 지표들은 다음과 같이 정의된다. 분류의 정확도를 나타내는 Precision(Sokolova and Lapalme, 2009)은 식 (5)와 같다.

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

실제 양성 표본 중에서 모델이 양성으로 분류한 비율을 나타내는 Recall(Sokolova and Lapalme, 2009)은 식 (6)과 같다.

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

Precision과 Recall의 조화평균인 F-measure(Sokolova and Lapalme, 2009)는 식 (7)과 같이 정의된다.

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

불균형 데이터에서 양쪽 클래스 성능 균형을 평가하기 위해 사용되는 G-mean(He *et al.*, 2009)은 식 (8)과 같이 TPR(True Positive Rate)과 TNR(True Negative Rate)의 기하평균으로 정의된다.

$$G-mean = \sqrt{TPR \times TNR} \tag{8}$$

마지막으로, 클래스 불균형에도 강건한 상관관계 지표인 Matthews Correlation Coefficient(Chicco and Jurman, 2020)는 식 (9)와 같이 TP, TN, FP, FN 모두를 활용하여 계산한다.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{9}$$

### 4.2 실험 결과

<Figure 3>은 4가지 RUS 비율(None, 0.1, 0.3, 0.5)에서 10회 반복한 각 분류기의 AUC 변화를 보여준다. 모든 비율에서 CatBoost가 Xgboost와 Random Forest를 앞서 가장 높은 AUC를 기록하여, 클래스 분포 조정 여부와 무관하게 우수한 성능을 유지하였다. 본 연구에서는 가장 우수한 성능을 보인 CatBoost를 분류기로 선정하고 결과를 제시한다.

<Table 5>는 RUS를 적용하지 않은 원본 데이터에서 임곗값 최적화 지표별로 도출한 최적 임곗값(Threshold)과 CatBoost의 성능(Metric)을 정리한 표다. 여기서 D는 기본 임곗값(Default Threshold), C는 데이터의 양성 사전 확률(Class Prior), NC는 제약 조건이 적용되지 않는 None Constraint를 의미한다. 기본 임곗값(D)을 적용하면 TPR은 0.0523, FNR은 0.9478로 매우 낮은 민감도(TPR)를 보였다. 반면, 'TPR ≥ TNR' 제약 아래에서 F-measure를 기준으로 임곗값을 최적화하면 TPR = 0.7537, TNR = 0.6308으로 민감도와 특이도가 모두 크게 향상되면서 균형이 크게 개선($|TPR-TNR| = 0.1229$)되었다. 그러나 제약을 포함하지 않는 F-measure NC 조건에서는 TPR = 0.2164로 많이 감소하고 TNR = 0.9647로 과도하게 상승($|TPR-TNR| = 0.7483$)하였다. 이는 제약 조건을 포함한 임곗값 탐색이 TPR과 TNR 간 균형을 개선함을 확인했다.

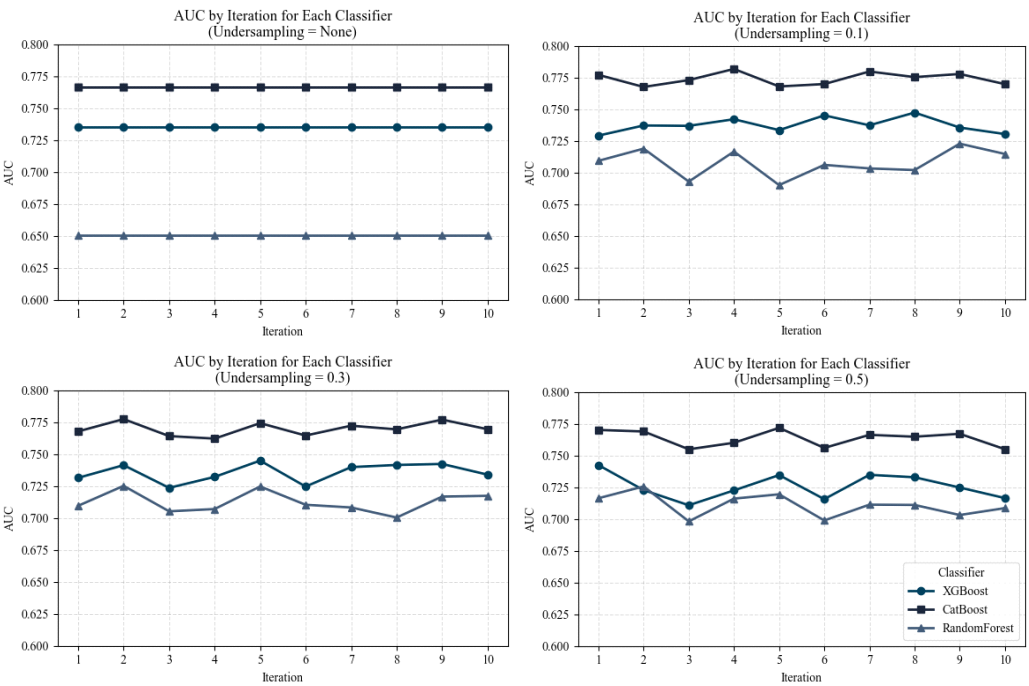



**Figure 3.** AUC by iteration across classifiers and undersampling ratios

<Table 6>은 RUS = 0.1 조건에서 임곗값 최적화 지표별로 도출한 최적 임곗값(Threshold)과 CatBoost의 성능(Metric)을 정리한 표다. 기본 임곗값(D)을 적용했을 때 TPR = 0.1194로 낮고 FNR = 0.8881로 낮은 민감도를 보였으나, RUS = 0.1 적용 전 원본 데이터 분포 대비 TPR이 상승하고 FNR이 감소하였다. 또한 F-measure NC를 기준으로 한 최적화 시 TPR = 0.4328, TNR = 0.8850을 기록하여, RUS를 미적용 한 실험에서 F-measure NC 기준 최적화 했을 때의 결과($|TPR-TNR| = 0.7483$)보다 민감도(TPR)와 특이도(TNR) 간 균형($|TPR-TNR| = 0.4522$)이 개선되었다.

<Table 7>은 RUS = 0.3 조건에서 임곗값 최적화 지표별로 도출한 최적 임곗값(Threshold)과CatBoost의 성능(Metric)을 보여준다. 기본 임곗값(D)을 적용했을 때 TPR = 0.2464, FNR = 0.7537로 낮은 민감도(TPR)를 보였으나, RUS = 0.1 적용 시에 비해 TPR이 상승하고 FNR이 감소하여 RUS 비율이 높아질수록 민감도를 개선함을 확인했다. 또한 F-measure NC를 기준으로 최적화 시 TPR = 0.7313, TNR = 0.7057을 기록하여, RUS

**Table 5.** CatBoost Classification Results (RUS = None)

| Objective | Threshold | Metric | | | |
|---|---|---|---|---|---|
| | | TPR | TNR | FPR | FNR |
| D | 0.500000 | 0.05223 | 0.99849 | 0.00150 | 0.94776 |
| C | 0.027957 | 0.67164 | 0.73093 | 0.26906 | 0.32835 |
| F-meas | 0.019745 | 0.75373 | 0.63076 | 0.36923 | 0.24626 |
| F-meas NC | 0.098879 | 0.21641 | 0.96467 | 0.03533 | 0.78358 |
| G-mean | 0.019249 | 0.77611 | 0.62667 | 0.37333 | 0.22388 |
| G-mean NC | 0.021144 | 0.73880 | 0.64907 | 0.35092 | 0.26119 |
| MCC | 0.019249 | 0.77611 | 0.62667 | 0.37333 | 0.22388 |
| MCC NC | 0.387941 | 0.08209 | 0.99741 | 0.00258 | 0.91791 |
| Precision | 0.020267 | 0.75373 | 0.63830 | 0.36169 | 0.24626 |
| Precision NC | 0.732190 | 0.02985 | 0.99935 | 0.00064 | 0.97014 |

**Table 6.** CatBoost Classification Results (RUS = 0.1)

| Objective | Threshold | Metric | | | |
|---|---|---|---|---|---|
| | | TPR | TNR | FPR | FNR |
| D | 0.500000 | 0.11194 | 0.99461 | 0.00538 | 0.88806 |
| C | 0.090909 | 0.70895 | 0.71133 | 0.28866 | 0.29104 |
| F-meas | 0.074580 | 0.76119 | 0.65079 | 0.34920 | 0.23880 |
| F-meas NC | 0.150570 | 0.43283 | 0.88496 | 0.11503 | 0.56716 |
| G-mean | 0.073316 | 0.76865 | 0.64713 | 0.35286 | 0.23134 |
| G-mean NC | 0.075560 | 0.75373 | 0.65424 | 0.34575 | 0.24626 |
| MCC | 0.072679 | 0.76865 | 0.64196 | 0.35803 | 0.23134 |
| MCC NC | 0.211643 | 0.32089 | 0.93989 | 0.06010 | 0.67910 |
| Precision | 0.076708 | 0.75373 | 0.65769 | 0.34230 | 0.24626 |
| Precision NC | 0.857965 | 0.02985 | 0.99760 | 0.00138 | 0.97837 |

**Table 7.** CatBoost Classification Results (RUS = 0.3)

| Objective | Threshold | Metric | | | |
|---|---|---|---|---|---|
| | | TPR | TNR | FPR | FNR |
| D | 0.500000 | 0.24626 | 0.94032 | 0.05967 | 0.75373 |
| C | 0.230769 | 0.76119 | 0.68268 | 0.31732 | 0.23880 |
| F-meas | 0.207675 | 0.79104 | 0.63334 | 0.36665 | 0.20895 |
| F-meas NC | 0.244722 | 0.73134 | 0.70573 | 0.29427 | 0.26865 |
| G-mean | 0.205889 | 0.79104 | 0.62968 | 0.37031 | 0.20895 |
| G-mean NC | 0.232998 | 0.75373 | 0.68569 | 0.31430 | 0.24626 |
| MCC | 0.200424 | 0.79850 | 0.61611 | 0.38388 | 0.20149 |
| MCC NC | 0.255239 | 0.70895 | 0.71887 | 0.28112 | 0.29104 |
| Precision | 0.213872 | 0.79104 | 0.64562 | 0.35437 | 0.20895 |
| Precision NC | 0.873121 | 0.03731 | 0.99827 | 0.00172 | 0.96268 |

= 0.1 조건에서의 F-measure NC를 기준으로 최적화했을 때의 결과($|TPR-TNR| = 0.4522$)보다 균형($|TPR-TNR| = 0.0256$)이 개선되었다. 이에 따라 RUS 비율을 높일수록 민감도(TPR)와 특이도(TNR)간 균형을 개선할 수 있다.

<Table 8>은 RUS = 0.5 조건에서 임곗값 최적화 지표별로 도출한 최적 임곗값(Threshold)과 CatBoost의 성능(Metric)을 정리하였다. 기본 임곗값(D)을 기준으로 최적화 시 TPR = 0.4179, FNR = 0.5821을 기록하며 RUS = 0.3 적용 시에 비해 TPR이 상승하고 FNR이 감소하였다. 하지만 F-measure NC 기준 최적화 시 TPR = 0.7836, TNR = 0.6206로, 앞선 RUS = 0.3 조건에서의 결과($|TPR-TNR| = 0.0256$) 대비 민감도와 특이도 간 균형($|TPR-TNR| = 0.1630$)이 개선되지 않았다. 이는 RUS 비율을 0.3 이상으로 높여도 추가적인 균형 향상이 없음을 알 수 있다.

추가적으로, 각 RUS 조건에서의 임곗값 및 성능 곡선을 시각화하였다. 분석 결과, RUS = None에서는 G-Mean, MCC, F1 지표가 0.5~0.8 구간에서, RUS = 0.1에서는 0.8~0.95 구간에서 변화 폭이 거의 없는 정체 구간이 나타남을 <Figure 4>에서 확인하였다. 다만 정체 구간은 최적 임곗값 산출에는 큰 영향을 미치지 않았으며, 모든 지표의 최댓값은 정체 구간 밖에서 나

**Table 8.** CatBoost Classification Results (RUS = 0.5)

| Objective | Threshold | Metric | | | |
|---|---|---|---|---|---|
| | | TPR | TNR | FPR | FNR |
| D | 0.500000 | 0.41791 | 0.86277 | 0.13722 | 0.58209 |
| C | 0.333333 | 0.70895 | 0.70185 | 0.29814 | 0.29104 |
| F-meas | 0.255192 | 0.80597 | 0.58035 | 0.41964 | 0.19403 |
| F-meas NC | 0.273750 | 0.78358 | 0.62063 | 0.37936 | 0.21641 |
| G-mean | 0.296733 | 0.76119 | 0.65596 | 0.34403 | 0.23880 |
| G-mean NC | 0.324218 | 0.71641 | 0.69474 | 0.30525 | 0.28358 |
| MCC | 0.254025 | 0.80597 | 0.57949 | 0.42050 | 0.19403 |
| MCC NC | 0.439754 | 0.53731 | 0.81710 | 0.18289 | 0.46268 |
| Precision | 0.297161 | 0.76119 | 0.65639 | 0.34360 | 0.23880 |
| Precision NC | 0.899996 | 0.05223 | 0.99526 | 0.00473 | 0.94776 |

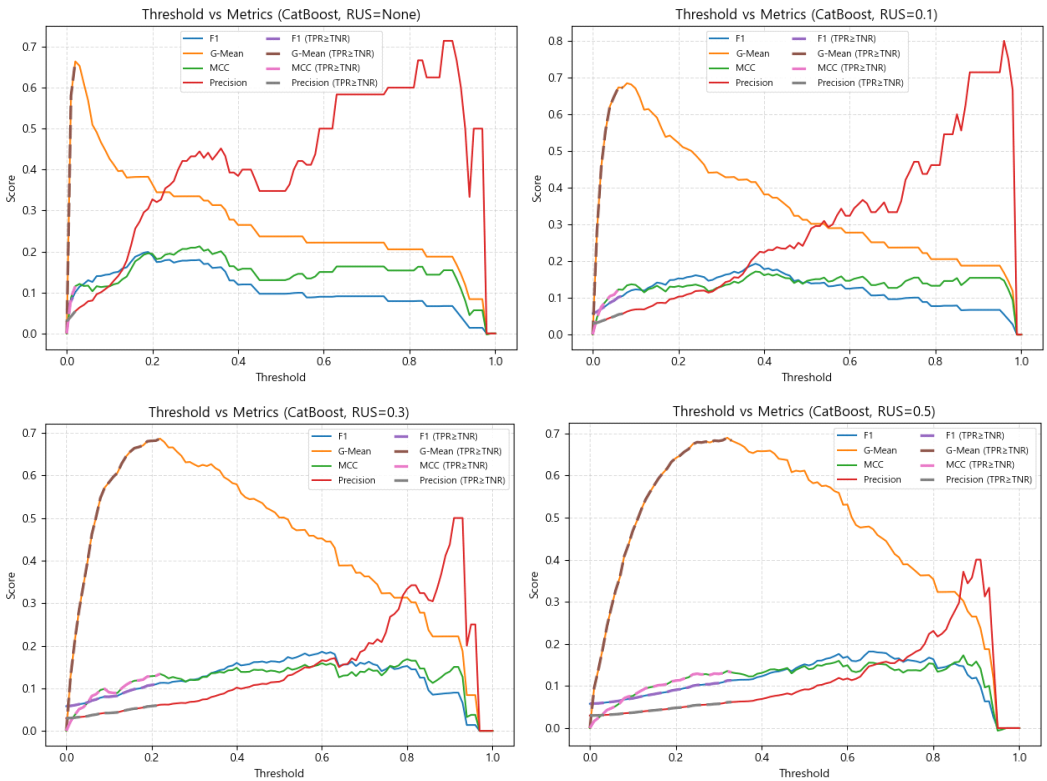



**Figure 4.** Threshold-Performance Metric Curves under Different RUS Ratios

타났다. 점선으로 표시된 결과는 제약조건 하에서 정체 구간의 존재 여부를 확인하기 위함이며, 제약조건 하에서는 정체 구간이 나타나지 않았고, 최적 임곗값 산출에도 영향을 미치지 않음을 확인하였다.

### 4.3 가설 검증

4.2절에서의 실험 결과를 바탕으로 앞서 제시한 가설들을 검증한다.

가설 1의 검증을 위해 모든 실험 조건(원본, RUS=0.1, RUS=0.3, RUS=0.5)에서 기본 임곗값(D)을 적용한 결과, TPR이 낮고 FNR이 높아 소수 클래스 검출에는 한계가 있다. 예를 들어, 원본 데이터에서는 TPR = 0.0522로 전체 양성 표본의 95% 이상을 검출하지 못하였다. 따라서 불균형이 심한 용접 결함 분류 문제에서는 기본값 0.5 임곗값이 부적합함을 확인했다.

가설 2 검증에서는 RUS 비율을 0.1, 0.3, 0.5로 증가시키며 네 가지 지표(F-measure, G-Mean, MCC, Precision) 기준 최적 임곗값 변화를 분석했다. 그 결과 모든 실험에서 최적 임곗값이 상승했으며, RUS = 0.5 조건의 F-measure 기준 최적 임곗값은 약 0.2551로, 원본(0.0197) 대비 13배 이상 증가하였다. 이로써 소수 클래스 비율이 높아질수록 임곗값이 높아진다는 것을 입증하였다.

가설 3 검증을 위해 원본 데이터에서 F-measure 기준 최적화 시 제약 적용 전과 후의 TPR과 TNR 값을 비교하였다. 제약을 적용하지 않은 경우(F-measure NC) TPR은 0.2164, TNR은 0.9647로 두 값의 차이는 0.7483이지만, 제약을 적용한 후(F-measure)에는 TPR이 0.7537, TNR이 0.6308로 차이가 0.1229로 크게 줄었다. 이로써 TPR $\geq$ TNR 제약이 민감도와 특이도 간의 불균형을 완화함을 검증했다.

## 5. 결 론

본 연구는 3개년의 현장 기록을 통합한 배관 용접 결함 데이터세트를 활용해, 클래스 불균형을 해소하면서 용접 결함을 분류할 수 있는 방법론을 제시한다. 클래스 불균형을 해소하기 위해 Random Undersampling(RUS)을 적용하고, 임곗값 최적화(Threshold Opimization) 방법을 도입하였다. 임곗값 최적화(Threshold Optimization)는 F-measure, G-mean, MCC, Precision 네 가지 지표를 최적화 성능 지표로 사용하였으며, 분류기 성능 평가는 AUC, TPR, TNR, FPR, FNR 다섯 지표를 기준으로 분석하였다. 또한 최적 임곗값 탐색 시 'TPR $\geq$ TNR'제약을 적용하여 민감도(TPR)와 특이도(TNR)의 균형을 개선하였다.

모든 실험에서 CatBoost가 가장 높은 AUC로 일관되게 우수한 성능을 보였다. RUS 비율이 증가할수록 최적 임곗값이 상승하였고, 기본 임곗값 0.5를 기준으로 TPR이 상승하고 FNR이 감소하여 민감도와 특이도가 개선되었다. 또한 임곗값 최적화에서 'TPR $\geq$ TNR' 제약을 적용했을 때 민감도(TPR)와 특이도(TNR) 간 편향이 감소하였다.

이를 통해 기본 임곗값 0.5는 모든 RUS 수준에서 양성 클래스 검출 성능이 낮아 클래스 불균형 환경에서 부적합하다는 결론을 도출할 수 있었다. 또한 양성 클래스 사전 확률(양성 클래스 비율)이 높아질수록 최적 임곗값이 상승하는 결과를 통해 데이터를 조정하는 방법(RUS)과 모델의 학습 및 분류 과정을 개선하는 방법(임곗값 최적화)을 병행하면 클래스 불균형 상황에서도 TPR과 TNR 간 균형 잡힌 용접 결함 분류가 가능하다는 것을 보여준다.

하지만 본 연구는 다음과 같은 한계점을 가진다. 클래스 불균형을 해소하기 위한 데이터를 조정하는 방법으로 RUS만을 사용하여 다른 방법들과 비교하지 않았고, 트리 기반 모델에 한정하였으므로 딥러닝 모델과 같은 최신 모델을 검토하지 못한 한계점이 있다. 따라서 향후 연구에서는 다양한 데이터 조정 방법과의 비교와 딥러닝 기반 모델 등의 적용을 통해 모델의 확장성과 일반화를 고려한 연구를 수행하고자 한다.

## 참고문헌


Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., and Siddique, Z. (2021), Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance, *Technologies*, **9**(3), 52.

Apeh, S. E., Kuye, S. I., Adetunji, O. R., and Anyanwu, B. U.(2023), A review of some welding parameters and their effects on the heat-affected zone of mild steel plate, *Nigerian Journal of Materials Science and Engineering*, **13**(1), 1-11.

Araf, I., Idri, A., and Chairi, I. (2024), Cost-sensitive learning for imbalanced medical data: A review, *Artificial Intelligence Review*, **57**, 80.

Arafa, A., El-Fishawy, N., Badawy, M., and Radad, M. (2022), RN-SMOTE: Reduced Noise SMOTE Based on DBSCAN for Enhancing Imbalanced Data Classification, *Journal of King Saud University - Computer and Information Sciences*, 34(10), 5059-5074.

Breiman, L. (2001), Random Forests, *Machine Learning*, **45**(1), 5-32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002), SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research*, **16**, 321-357.

Chen, T. and Guestrin, C. (2016), XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Cheng, H., Kang, L., Pang, J., Xue, B., Du, D., and Chang, B. (2021), Effect of the welding position on weld quality when laser welding Inconel 617 Ni-based superalloy, *Optics and Laser Technology*, **139**, 106962.

Chicco, D. and Jurman, G. (2020), The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC Genomics*, **21**, Article 6.

Gao, J., Liang, M., Wang, J., Zha, S., Yang, A., and Lan, H. (2025),

Analysis of residual stress of butt fusion joints for polyethylene gas pipes, *Polymers*, **17**(10), 1388-1406.

González, S., García, S., Del Ser, J., Rokach, L., and Herrera, F. (2020), A Practical Tutorial on Bagging and Boosting Based Ensembles for Machine Learning: Algorithms, Software Tools, Performance Study, *Practical Perspectives and Opportunities, Information Fusion*, **64**, 205-237.

Gupta, A., Nagarajan, V., and Ravi, R. (2017), Approximation Algorithms for Optimal Decision Trees and Adaptive TSP Problems, *Mathematics of Operations Research*, **42**(3), 876-896.

Hahn, Y., Maack, R., Tercan, H., Meisen, T., Purrio, M., Buchholz, G., and Angerhausen, M.(2023), Towards a Deep Learning-Based Online Quality Prediction System for Welding Processes, *arXiv preprint* arXiv:2310.12632.

Hancock, J. T. and Khoshgoftaar, T. M. (2020), Survey on categorical data for neural networks, *Journal of Big Data*, **7**, 28.

Hancock, J., Johnson, J. M., and Khoshgoftaar, T. M. (2022), A Comparative Approach to Threshold Optimization for Classifying Imbalanced Data, *Proceedings of the IEEE International Conference on Collaborative Internet Computing*, 135-142.

Hancock, J., Khoshgoftaar, T. M., and Johnson, J. M. (2022), The Effects of Random Undersampling for Big Data Medicare Fraud Detection, *Proceedings of the 2022 IEEE International Conference on Service-Oriented System Engineering (SOSE)*, 141-152.

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., and Bauder, R. A. (2019), Severely Imbalanced Big Data Challenges: Investigating Data Sampling Approaches, *Journal of Big Data*, **6**(1), 107.

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

Hou, W., Wei, Y., Jin, Y., and Zhu, C. (2019), Deep features based on a DCNN model for classifying imbalanced weld flaw types, *Measurement*, **131**, 482-489.

Johnson, J. M. and Khoshgoftaar, T. M. (2021), Output Thresholding for Ensemble Learners and Imbalanced Big Data, *Proceedings of the 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence*, 1449-1454.

Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., and Hussain, S. (2023), Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks, *Applied Sciences*, **13**, 4006.

Kini, S. M., Devidas, B., Pai, S. N., Kolekar, S., Pai, V., and Balasubramani, R. (2022), Use of Machine Learning and Random Oversampling in Stroke Prediction, *Proceedings of the International Conference on Artificial Intelligence and Data Engineering (AIDE 2022)*, 331-337.

Kumar, V., Albert, S. K., Chandrasekhar, N., and Jayapandian, J. (2018), Evaluation of welding skill using probability density distributions and neural network analysis, *Measurement*, 116, 114-121.

Leevy, J. L., Johnson, J. M., Hancock, J., and Khoshgoftaar, T. M. (2023), Threshold Optimization and Random Undersampling for Imbalanced Credit Card Data, *Journal of Big Data*, **10**, 58.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., and Seliya, N. (2018), A Survey on Addressing High-Class Imbalance in Big Data, *Journal of Big Data*, **5**(1), 42.

Lemaître, G., Nogueira, F., and Aridas, C. K.(2017), Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, *Journal of Machine Learning Research*, **18**(1), 559-563.

Liu, J., Cheng, Y., Jing, X., Liu, X., and Chen, Y.(2024), Prediction and Optimization Method for Welding Quality of Components in Ship Construction, *Scientific Reports*, **14**, 9353.

Moinuddin, S. Q., Hameed, S. S., Dewangan, A. K., Kumar, K. R., and Kumari, A. S. (2021), A study on weld defects classification in gas metal arc welding process using machine learning techniques, *Materials Today: Proceedings*, **43**, 623-628.

Muthmann, E., Kaluza, W., Liedtke, M., and Scheller, W. (2009), Induction bends in material grade X80: Experience from more than 15 years, *Proceedings of the Pipeline Technology Conference*, Ostend, 12-14 October 2009; Paper no. Ostend2009-076.

Park, J.-K., An, W.-H., and Kang, D.-J. (2019), Convolutional neural network based surface inspection system for non-patterned welding defects, *International Journal of Precision Engineering and Manufacturing*, **20**(3), 363-374.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.,

Vanderplas, J., Passos, A., and Cournapeau, D. (2011), Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, **12**, 2825-2830.

Pradipta, G. A., Sanjaya, I. N. H., Wardoyo, R., Musdholifah, A., and Ismail, M. (2021), SMOTE for handling imbalanced data problem: a review, *Proceedings of the 2021 Sixth International Conference on Informatics and Computing (ICIC)*, 1-6.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018), CatBoost: Unbiased Boosting with Categorical Features, *Advances in Neural Information Processing Systems*, **31**, 6638-6648.

Shaloo, M., Schnall, M., Klein, T., Huber, N., and Reitinger, B. (2022), A review of non-destructive testing (NDT) techniques for defect detection: Application to fusion welding and future wire arc additive manufacturing processes, *Materials*, **15**(10), 3697-3712.

Sokolova, M. and Lapalme, G. (2009), A systematic analysis of performance measures for classification tasks, *Information Processing & Management*, **45**(4), 427-437.

Tian, J., Tsai, P.-W., Zhang, K., Cai, X., Xiao, H., Yu, K., Zhao, W., and Chen, J. (2024), Synergetic Focal Loss for Imbalanced Classification in Federated XGBoost, *IEEE Transactions on Artificial Intelligence*, **5**(2), 647-657.

Vasan, V., Sridharan, N. V., Balasundaram, R. J., and Vaithiyanathan, S. (2024), Ensemble-Based Deep Learning Model for Welding Defect Detection and Classification, *Engineering Applications of Artificial Intelligence*, **136**, 108961.

Wang, G.-Q., Zhang, C.-Z., Chen, M.-S., Lin, Y.-C., Tan, X.-H., Liang, P., Kang, Y.-X., Zeng, W.-D., and Wang, Q. (2023), Yolo-MSAPF: Multiscale alignment fusion with parallel feature filtering model for high accuracy weld defect detection, *IEEE Transactions on Instrumentation and Measurement*, **72**, 5022914.

Wang, Z., Xu, Q., Yang, Z., He, Y., Cao, X., and Huang, Q. (2023), A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning, *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

Wongvorachan, T., He, S., and Bulut, O. (2023), A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining, *Information*, **14**(1), 54.

Yang, D., Cui, Y., Yu, Z., and Yuan, H. (2021), Deep learning based steel pipe weld defect detection, *Applied Artificial Intelligence*, **35**(15), 1237-1249.

Zhang, Z., Wen, G., and Chen, S. (2019), Weld image deep learning-

based on-line defects detection using convolutional neural networks for Al alloy in robotic arc welding, *Journal of Manufacturing Processes*, **45**, 208-216.

## 저자소개

**한민기:** 부산대학교 산업공학과 학사과정에 재학 중이다. 연구 분야는 데이터마이닝, 빅데이터 분석, 인공지능이다.

**김도희:** 부산대학교 산업공학과에서 2019년 학사학위를, 2024년 박사학위를 취득하였다. 이후 2025년 11월까지 부산대학교 인간중심 탄소중립 글로벌공급망 연구센터에서 연수연구원과 연구교수로 재직하였다. 2025년 12월부터는 국립창원대학교 인공지능융합공학과에서 조교수로 재직 중이다. 주요 연구 관심 분야는 시계열 분석, 인공지능, 딥러닝이다.

**배혜림:** 서울대학교에서 1996년에 학사, 1996년 석사, 2002년 박사학위를 취득하였고, 2002년부터 2003년까지 삼성카드에서 근무했으며, 2005년부터 부산대학교 산업공학과 교수를 역임하고 2024년부터 부산대학교 데이터사이언스전문대학원 교수로 재직 중이다. 관심 분야는 정보시스템 설계, 클라우드 컴퓨팅, 비즈니스 프로세스 마이닝, 항만 물류, 인공지능이다.