https://doi.org/10.7735/ksmte.2026.35.5.404

J. Korean Soc. Manuf. Technol. Eng.
ISSN 2508-5107(Online)

Check for updates

용접검사 반복영상에서 평가단위와 프레임 수에 따른 분류성능 변화

윤준석[a], 원홍인[a,*]

Effects of Evaluation Unit and Frame Count on Classification Performance in Repeated Welding Inspection Images

Jun-Seok Yun[a], Hong-In Won[a,*]

[a] Manufacturing AI Research Center, Korea Institute of Industrial Technology (KITECH)

ARTICLE INFO

Article history:
Received 24 August 2026
Revised 9 September 2026
Accepted 18 September 2026

Keywords:
Manufacturing inspection
Welding inspection
Repeated acquisition
Evaluation unit
Grouped split
Data leakage

ABSTRACT

Evaluating manufacturing vision requires strict isolation of repeated acquisitions between training and test data. We analyzed 1,120 nonpublic welding inspection images from two quality classes and 140 acquisition groups. After removing exact duplicates, we varied unique frames per group from one to five and compared image-random, acquisition-group, and paired-group five-fold validations across 30 assignments. Same-group exposure under image-random splitting reached 99.9%. Consequently, the balanced accuracy of HOG-linear SVM increased from 0.376 to 0.762 under image-random splitting but decreased from 0.409 to 0.288 under acquisition-group splitting, while paired-group performance remained at ~0.55. A compact convolutional baseline showed a much smaller but directionally consistent trend, with balanced accuracies of 0.524 (image random) and 0.484 (acquisition group). Decision-margin directions changed with the split unit, thus rendering model superiority unstable. Therefore, repeated frames amplify evaluation bias without adding independent cases. Performance reports should quantify exposure and align split units with intended deployment cases.

1. 서 론

최근 제조 현장에서는 검사 공정의 무인화를 목표로 딥러닝 기반 외관검사 기술의 도입이 활발하게 이루어지고 있다. 이러한 연구에서는 분류정확도와 처리속도를 주요 성능지표로 사용하며, MLCC 외관검사 장비 연구에서는 검사정확도와 처리속도를 함께 검증하였고[1], 자동차 부품 결함분류 연구에서는 영상 취득환경에 따른 CNN 분류성능을 분석하였다[2]. 또한 표면결함 검출성능은 조명, 배경, 소재 및 결함형상과 같은 영상 취득조건에 의해서도 크게 좌우되는 것으로 보고되었다[3]. 그러나 이러한 성능지표가 실제 현장 적용 가능성을 나타내기 위해서는, 성능 산출에 사용되는 테스트 데이터가 AI 모델이 새로 판정하게 될 검사사례를 대표할 수 있어야 한다.

특히 용접검사와 같이 한 부품이나 고정된 촬영위치에서 여러 프레임을 반복 저장하는 검사 공정에서는, 저장된 영상파일 수가 서로 독립된 검사사례 수보다 많아지게 된다. 이때 동일한 검사사례에서 취득된 영상의 집합을 촬영묶음이라 하며, 하나의 촬영묶음에 속한 영상이 학습 데이터와 테스트 데이터에 나뉘어 배정될 경우 문제가 발생한다. AI 모델이 검사 대상의 품질특성이 아니라 학습 과정에서 이미 관찰한 형상이나 배경을 다시 식별하여 판정

* Corresponding author. Tel.: +82-41-589-8190
E-mail address: luvhayym@kitech.re.kr (Hong-In Won).

할 수 있기 때문이다. 이와 같이 학습데이터와 테스트데이터가 정보를 공유하는 현상은 데이터 누수로 알려져 있으며, 보고되는 성능을 실제보다 높게 만드는 주요 원인으로 지적되어 왔다[4-6]. 이러한 이유로 계층 구조나 시공간적 종속성을 갖는 데이터에서는 생성단위를 보존한 교차검증이 요구된다[7].

중복 저장된 영상은 파일 해시값 비교로 식별·제거할 수 있지만, 같은 촬영묶음에서 얻은 서로 다른 프레임이 학습데이터와 테스트데이터에 나뉘는 문제는 이 절차만으로 해결되지 않는다. 이에 따라 분할 전에 중복과 생성단위의 중첩을 점검하고, 동일한 생성단위가 서로 다른 폴드에 배정되지 않도록 평가하는 절차가 제안되어 왔다[8,9]. 반복 촬영자료의 성능을 해석하려면 중복 저장영상의 제거 여부와 학습·테스트 분할에 사용한 평가단위를 함께 살펴볼 필요가 있다.

이에 본 연구는 외부에서 제공된 비공개 용접검사 영상을 대상으로 중복 저장영상, 촬영묶음, 서로 다른 품질등급 간의 파일기반 대응묶음 및 묶음당 고유 프레임 수를 계량하고, 이 구조를 달리한 분할이 성능 추정에 미치는 영향을 분석하였다. 학습·테스트 중첩 노출률을 산출하고, HOG 기반 선형 SVM과 저해상도 픽셀 기반 RBF-SVM을 기준모델로 사용하였으며, 표현학습 모델에서도 평가단위의 영향을 확인하기 위해 소형 CNN을 동일한 반복 교차검증 절차로 평가하였다. 또한 촬영묶음당 고유 프레임 수를 1~5장으로 변화시키고 대응 촬영묶음 분할을 반복하여, 프레임 수와 그룹 배정에 따른 성능 추정치 및 모델 순위의 변동을 비교하였다.

2. 데이터 및 평가설계

2.1 용접검사 영상 데이터

본 연구에서 사용한 데이터는 외부에서 제공된 비공개 용접검사 영상 1,120장으로, 제공 기록에는 양품·불량의 이진 라벨과 파일 구조가 포함되어 있다. 파일명 상위식별자를 기준으로 140개의 촬영묶음을 구성하였으며 각 묶음에는 8장의 영상이 포함된다. 서로 다른 품질등급에서 같은 상위식별자를 공유하는 두 촬영묶음은 본 논문에서 '파일기반 대응묶음'으로 정의하였다. 다만 제공 기록에는 라벨의 세부 판정기준과 실제 부품·촬영시점·로트 식별자가 포함되어 있지 않으므로, 이 대응관계는 물리적으로 동일한 검사대상을 뜻하지 않고 파일구조에서 확인되는 분석단위를 뜻한다.

또한 동일한 촬영묶음 내에 완전히 동일한 영상이 포함되어 있는지 확인하기 위하여, 각 영상 파일의 내용에 대한 해시값을 산출하고 이를 상호 비교하였다. 그 결과 전체 1,120장 가운데 420장이 중복 저장영상으로 판정되었으며, 이를 제거하면 모든 촬영묶음에 각각 5장씩 총 700장의 고유 영상이 남는다. 즉 촬영묶음의 수는 반복영상의 제거 여부와 무관하게 140개로 일정하게 유지된다. Table 1에 데이터 구성을 정리하였다.

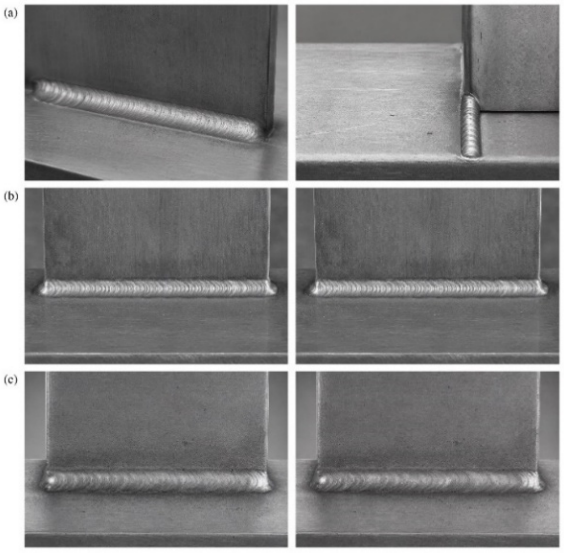


Fig. 1 Illustrative reconstruction of image relationships in the welding inspection data: (a) distinct acquisition cases, (b) an exact duplicate within an acquisition group, and (c) visually corresponding groups across two anonymous quality classes

Table 1 Configuration of the welding inspection image data

Item	Value
Quality classes	2
Total images	1,120
Acquisition groups	140
Images per acquisition group	8
Images after exact-repeat removal	700
Exactly repeated images	420
Cross-class paired-group candidates	70 pairs

원영상과 파일명, 업체·현장·장비의 식별정보는 계약 조건과 제3자 권리를 고려하여 공개하지 않고, 분석 절차와 익명화한 집계결과만 제시하였다. Fig. 1은 데이터에서 확인된 관계를 설명하기 위한 재구성도로, 서로 다른 촬영사례, 동일 촬영묶음 안의 중복 저장영상, 두 익명 품질등급 사이의 파일기반 대응묶음을 구분하여 나타낸다.

2.2 촬영묶음 대응구조 분석

서로 다른 품질등급으로 분류된 촬영묶음 사이에도 종속관계가 존재하는지를 확인하기 위하여, 파일명에서 동일한 상위 식별자를 공유하는 촬영묶음의 쌍을 조사하였다. 그 결과 두 품질등급에 걸쳐 대응 촬영묶음 70쌍이 존재하는 것으로 나타났으며, 이러한 대응관계가 파일명뿐 아니라 영상내용에서도 확인되는지를 픽셀 특징과 HOG 특징을 이용하여 검증하였다.

먼저 모든 영상을 저해상도 회색조로 정규화한 뒤, 픽셀값에 대해서는 정규화 상호상관계수를, HOG 특징에 대해서는 코사인 유사도를 각각 계산하였다. 이때 용접부의 촬영 방향이 좌우로 달라질 수 있는 점을 고려하여, 원영상과 수평반전 영상에서 얻은 두 유사도 중 큰 값을 해당 쌍의 유사도로 사용하였다. 이어서 한 품질등급의 촬영묶음을 기준으로 다른 등급의 후보 70개를 검색하고, 파일명으로 지정된 대응 촬영묶음이 1순위 및 상위 5개 후보에 포함되는 비율을 산출하였다. 또한 대응 촬영묶음과 비대응 촬영묶음의 평균유사도 차이가 우연에 의한 것인지 확인하기 위하여 10,000회의 무작위 순열 기반 검정을 수행하였다.

다만 본 절에서 다루는 대응관계는 영상의 형상과 구도에서 관찰되는 유사성에 근거한 것이며, 실제 부품이나 촬영 주기의 동일성을 의미하지는 않는다.

2.3 반복영상 처리와 분할조건

반복영상의 처리 방식과 평가단위가 성능에 미치는 영향을 분리하여 관찰하기 위하여, 반복영상의 보존 여부와 분할단위를 교차한 2×2 형태의 네 가지 조건을 구성하였다. 각 조건은 다음과 같다.

1)영상 무작위 분할은 전체 1,120장을 품질등급 비율에 따라 나눈다.

2) 반복 제거 후 무작위 분할은 고유 영상 700장을 같은 방식으로 나눈다.

3) 촬영묶음 분할은 1,120장을 사용하되 동일한 촬영묶음이 학습데이터와 테스트데이터로 분리되지 않도록 한다.

4) 반복 제거 후 촬영묶음 분할은 고유 영상 700장에 촬영묶음 분할을 적용한다.

네 조건 모두에 대하여 서로 다른 난수 초기값 10개와 5겹 교차검증을 사용하였으며, 영상 무작위 조건에는 계층화 분할을, 촬영묶음 조건에는 계층화 그룹분할을 각각 적용하였다. 한편 2.2절에서 확인된 대응 촬영묶음의 영향을 통제하기 위한 민감도 분석으로, 서로 다른 품질등급의 대응 촬영묶음을 하나의 단위로 묶은 조건을 추가하였다. 이 조건에서는 70개의 대응그룹을 무작위로 섞어 크기가 같은 5개 폴드에 배정하는 절차를 30회 반복하였다.

평가설계의 영향을 서로 다른 특징표현에서 확인하기 위해 회색조 HOG 특징을 입력으로 하는 선형 SVM과 64×48 회색조 픽셀값을 입력으로 하는 RBF-SVM을 기준모델로 사용하였다. HOG는 128×96 회색조 영상에서 9방향 빈, 8×8 셀, 2×2 블록, L2-Hys 정규화로 추출하였다. HOG-SVM은 학습 폴드에서만 적합한 StandardScaler와 선형 커널(C=1.0)을, 저해상도 SVM은 RBF 커널(C=1.0, gamma=scale)을 사용하였으며 두 분류기 모두 balanced class weight를 적용하였다.

표현학습 기반의 보조 비교모델은 64×48 회색조 입력과 16·32 채널의 두 합성곱 블록으로 구성한 소형 CNN으로 설정하였다. AdamW 최적화기(학습률 0.001, 가중치 감쇠 0.0001), 배치 크기 64, 30 epoch를 사용하였다. 모든 모델의 각 조건은 10개 난수 초기값에서 평가하였고, 각 폴드의 영상별 예측점수를 촬영묶음 안에서 평균하여 촬영묶음 수준 성능을 계산하였다. 주 평가지표는 두 품질등급의 재현율을 평균한 균형정확도(balanced accuracy, BA)이며 그 정의는 식 (1)과 같다.

$$\mathrm{BA} = \frac{1}{2}\left[\frac{\mathrm{TP1}}{(\mathrm{TP1}+\mathrm{FN1})} + \frac{\mathrm{TP2}}{(\mathrm{TP2}+\mathrm{FN2})}\right] \tag{1}$$

여기서 TP1과 FN1은 첫 번째 품질등급에 속하는 촬영묶음 중 해당 등급으로 옳게 분류된 수와 잘못 분류된 수를 나타내며, TP2와 FN2는 두 번째 품질등급에 대한 동일한 값이다. 대괄호 안의 두 항은 각 품질등급의 재현율에 해당한다. 보조지표로는 AUROC를 함께 산출하였다. 이하의 모든 결과는 서로 다른 난수 초기값 10개에 대한 평균과 표준편차로 보고하며, 대응 촬영묶음 조건의 경우 30회의 무작위 그룹 배정에 대한 평균과 표준편차를 사용하였다.

대응 촬영묶음 조건의 경우 두 SVM 기준모델은 30회의 무작위 그룹 배정에 대한 평균과 표준편차를 사용하였고, 소형 CNN은 제공된 파일구조에서 정해지는 고정 대응묶음 배정에 10개 난수 초기값을 적용하여 산출하였다. 또한 소형 CNN은 반복 제거 후 촬영묶음 분할을 제외한 네 개 조건에서 평가하였다.

2.4 중첩 노출률과 강건성 분석

분할조건에 따라 학습 과정에서 노출되는 정보량이 어떻게 달라지는지를 정량화하기 위하여, 각 폴드의 테스트영상마다 학습데이터에 동일 해시, 동일 촬영묶음 또는 동일 대응그룹에 해당하는 영상이 존재하는지를 확인하였다. 이때 해당 관계가 존재하는 테스트영상의 비율을 중첩 노출률로 정의하였으며, 이 값은 모델이 학습 과정에서 동일한 영상내용이나 촬영맥락을 접할 수 있었던 정도를 나타낸다. 세 가지 관계는 완전히 동일한 영상, 동일한 촬영묶음, 동일한 대응그룹의 순으로 포괄 범위가 넓어지므로, 이를 함께 산출하면 노출이 어느 수준에서 발생하는지를 구분할 수 있다.

또한 영상 무작위 분할과 촬영묶음 분할 사이의 BA 차이에 대해서는, 촬영묶음을 품질등급별로 재표집하는 5,000회의 부트스트랩을 통해 95% 신뢰구간을 산출하였다. 대응 촬영묶음 분석에서는 30개의 그룹 배정 각각에 대하여 BA와 두 모델의 성능 차이를 기록하였으며, 평균 순위뿐 아니라 각 모델이 더 높은 성능을 보인 횟수를 함께 보고함으로써 특정 배정에서 얻어진 결과에 결론이 좌우되지 않도록 하였다.

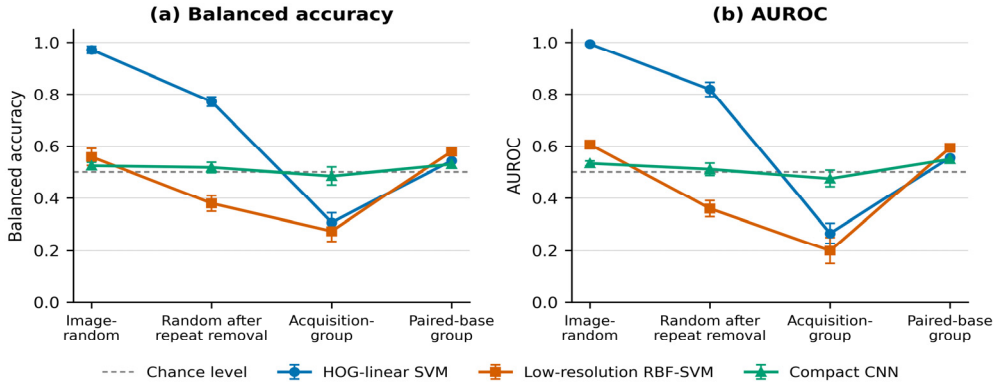


Fig. 2 Classification performance by evaluation split for the HOG-linear SVM, the low-resolution RBF-SVM, and the compact CNN: (a) balanced accuracy and (b) AUROC. Error bars denote standard deviations across ten seeds. The paired-group condition shown here uses the fixed grouping determined by the supplied file structure, whereas Table 3 reports means over 30 random paired-group assignments

2.5 프레임 수 민감도와 결정마진

촬영묶음당 저장되는 고유 프레임의 수가 평가편향에 미치는 영향을 확인하기 위하여, 중복 저장영상을 제거한 뒤 각 촬영묶음에 남은 고유 영상 5장 가운데 k장(k = 1~5)을 비복원 추출하여 표본을 구성하였다. 이렇게 얻은 동일한 표본에 대하여 영상 무작위 분할, 촬영묶음 분할 및 대응 촬영묶음 분할을 각각 적용함으로써, 분할단위 이외의 조건이 성능 차이에 개입하지 않도록 하였다. 표본추출과 5겹 분할은 30회 반복하였으며, 각 k에서 영상 무작위 분할과 촬영묶음 분할의 BA 차이를 산출하였다. 이때 촬영묶음의 수는 k의 값과 무관하게 140개로 유지되므로, 관찰되는 성능 변화는 독립 검사사례의 증가가 아니라 동일 사례의 프레임 반복에 의한 것으로 해석할 수 있다.

또한 분할조건에 따라 예측점수의 방향이 달라지는지를 비교하기 위하여, 테스트영상에 대한 SVM의 결정마진을 사용하였다. 각 폴드에서는 학습데이터 결정마진의 제곱평균제곱근으로 테스트점수를 나누어 척도를 정규화하되, 판정경계에 해당하는 0은 그대로 유지하였다. 이는 폴드마다 달라지는 마진의 절대 크기를 제거하고 점수의 방향과 상대적 크기만을 비교하기 위함이다. 이후 영상 단위의 점수를 촬영묶음 안에서 평균하고 이를 다시 30회 반복에 대하여 평균함으로써, 각 촬영묶음이 분포에 한 번씩만 반영되도록 하였다.

3. 결 과

3.1 반복영상과 촬영묶음 대응구조

전체 1,120장 가운데 420장은 동일한 촬영묶음 내에 존재하는 중복 저장영상으로 판정되었으며, 서로 다른 촬영묶음 사이에서는 완전 복제가 발견되지 않았다. 그 결과 140개의 모든 촬영묶음에서 8장 중 5장이 고유 영상으로 남았으며, 이는 반복 저장이 촬영묶음 내부에서만 균일하게 발생하였음을 의미한다.

Table 2 Training–test exposure according to the split condition

Split condition	Exact hash (%)	Same acquisition group (%)	Paired acquisition group (%)
Image-random	60.1±1.8	100.0	100.0
Random after repeat removal	0.0	100.0	100.0
Acquisition-group	0.0	0.0	68.1±8.4
Group after repeat removal	0.0	0.0	68.1±8.4
Paired acquisition-group	0.0	0.0	0.0

한편 파일명으로 지정된 대응 촬영묶음 70쌍을 비교한 결과, 픽셀 특징의 평균유사도는 대응그룹에서 0.861, 비대응그룹에서 0.222로 나타났으며, HOG 특징에서도 각각 0.805와 0.208로 동일한 경향을 보였다. 또한 1순위 검색 정확도는 픽셀 특징에서 95.7%, HOG 특징에서 97.1%였고, 5순위 이내 검색 정확도는 두 특징 모두 100%에 도달하였다. 두 특징에 대한 순열검정의 유의확률은 모두 0.001 미만으로, 파일명으로 지정한 대응관계가 영상의 형상과 구도에서도 동일하게 확인되었다.

3.2 분할조건별 학습·시험 중첩

Table 2는 각 분할조건에서 테스트영상이 학습데이터와 공유한 관계의 비율을 나타낸다. 영상 무작위 분할에서는 테스트영상의 60.1±1.8%가 학습데이터에 동일 해시를 갖는 영상을 가지고 있었으며, 모든 테스트영상이 동일 촬영묶음과 동일 대응그룹을 학습데이터와 공유하였다. 여기에서 반복영상을 제거하자 동일 해시에 의한 노출은 완전히 사라졌으나, 촬영묶음과 대응그룹 수준의 노출률은 여전히 100%로 유지되었다.

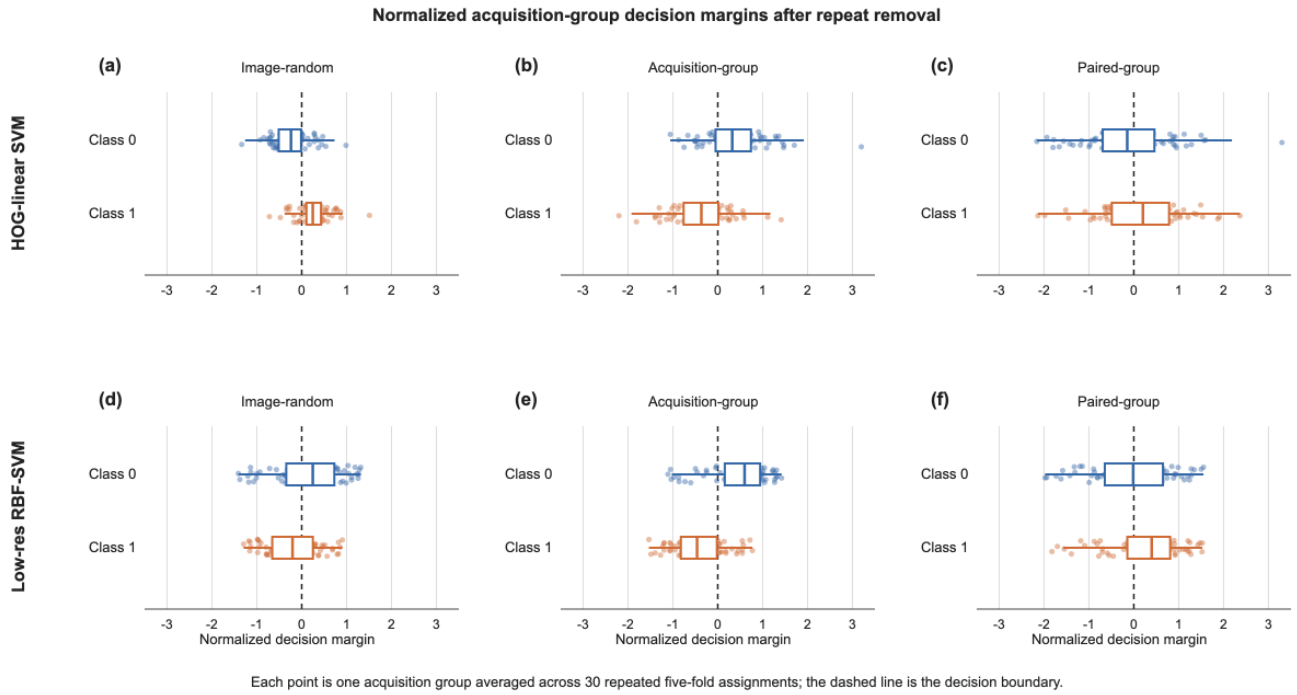


Fig. 3 Acquisition-group decision margins after exact-repeat removal under image-random, acquisition-group, and paired acquisition-group splitting. Margins were divided by the root mean square of training margins within each fitted fold, and each point denotes one acquisition group averaged across 30 repeated assignments

반면 촬영묶음 분할에서는 동일 해시와 동일 촬영묶음의 교차 노출이 발생하지 않았다. 다만 서로 다른 품질등급에 속한 대응 촬영묶음이 서로 다른 폴드에 배정되면서 대응그룹 노출률은 68.1±8.4%로 남았으며, 이는 촬영묶음을 분할단위로 사용하더라도 등급 간 대응관계까지는 통제되지 않음을 보여준다. 대응 촬영묶음을 하나의 분할단위로 묶은 조건에서는 노출률이 모두 0%로 나타났다.

3.3 분류성능과 모델 순위의 변동

Table 3과 Fig. 2에 분할조건별 분류결과를 제시하였다. Table 3은 두 SVM 기준모델의 결과이며, Fig. 2에는 소형 CNN을 포함한 세 모델의 결과를 함께 나타내었다. HOG-선형 SVM의 BA는 영상 무작위 분할에서 0.971로 가장 높았으나 촬영묶음 분할에서는 0.307까지 낮아졌으며, 중복 저장영상을 제거한 뒤에도 촬영묶음 분할의 BA는 0.313으로 거의 변화하지 않았다. 이는 영상 무작위 분할에서 관찰된 높은 성능이 중복 저장영상만으로는 설명되지 않음을 의미한다. RBF-SVM에서도 영상 무작위 분할에서 촬영묶음 분할로 갈수록 BA가 낮아지는 동일한 양상이 나타났으며, 반복영상 제거 여부에 따른 촬영묶음 분할의 BA 변화는 HOG 모델에서 0.006, RBF 모델에서 0.022로 두 모델 모두 작았다.

영상 무작위 분할과 촬영묶음 분할 사이의 BA 차이는 HOG 모델에서 0.664, RBF 모델에서 0.284였으며, 촬영묶음을 품질등급

Table 3 Classification performance according to the split condition

Model	Split condition	BA	AUROC
HOG-linear SVM	Image-random	0.971±0.014	0.994
HOG-linear SVM	Random after repeat removal	0.771±0.018	0.819
HOG-linear SVM	Acquisition-group	0.307±0.037	0.266
HOG-linear SVM	Group after repeat removal	0.313±0.043	0.267
HOG-linear SVM	Paired acquisition-group	0.551±0.024	0.569
Low-res RBF-SVM	Image-random	0.558±0.033	0.606
Low-res RBF-SVM	Random after repeat removal	0.379±0.029	0.360
Low-res RBF-SVM	Acquisition-group	0.274±0.039	0.200
Low-res RBF-SVM	Group after repeat removal	0.296±0.043	0.236
Low-res RBF-SVM	Paired acquisition-group	0.567±0.017	0.591

별로 재표집한 5,000회 부트스트랩의 95% 백분위 신뢰구간은 각각 0.593–0.750과 0.257–0.414로 모두 0을 포함하지 않았다. 또한 두 모델 모두 10개의 난수 초기값 전체에서 동일한 방향의 차이

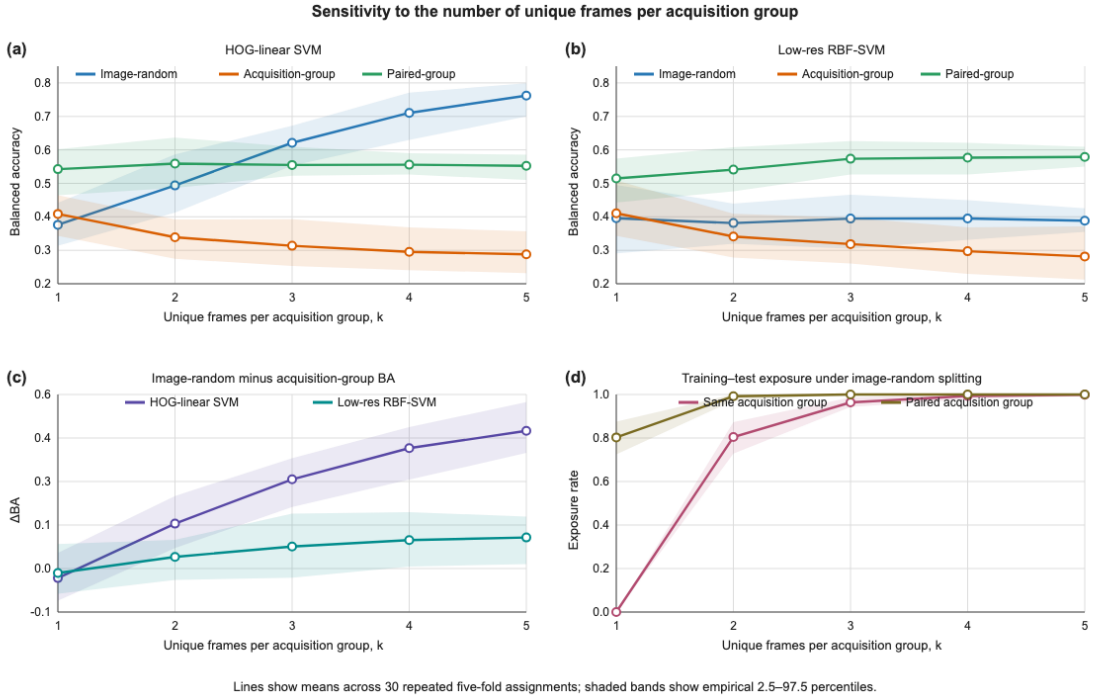


Fig. 4 Sensitivity to the number of unique frames per acquisition group: balanced accuracy for (a) HOG-linear SVM and (b) low-resolution RBF-SVM, (c) the image-random minus acquisition-group BA difference, and (d) training–test exposure under image-random splitting. Lines and bands denote means and empirical 2.5–97.5 percentiles across 30 repeated assignments

를 보였다.

촬영묶음 분할에서 HOG-SVM과 RBF-SVM의 AUROC는 반복영상 제거 여부와 관계없이 0.5보다 낮았고, 반복 제거 후 촬영묶음 분할에서는 품질등급 0의 평균 결정점수가 양수, 품질등급 1의 평균 결정점수가 음수로 나타났다. 반면 파일기반 대응묶음을 하나의 분할단위로 사용한 조건에서는 두 모델의 AUROC가 0.5를 상회하였다. 따라서 촬영묶음 분할에서 관찰된 점수 방향의 변화는 대응묶음이 학습·시험 세트에 배정되는 방식과 함께 나타난 현상으로 해석하였다.

대응 촬영묶음 분할을 30회 반복한 결과 평균 BA는 HOG 모델 0.551, RBF 모델 0.567이었다. 그러나 RBF 모델이 더 높은 배정이 18회, HOG 모델이 더 높은 배정이 10회, 두 모델이 동일한 배정이 2회로 나타났으며, 두 모델의 BA 차이는 평균 0.016, 2.5-97.5 백분위 범위는 -0.033에서 0.075였다. 즉 평균값에서는 RBF 모델이 우세하였으나 개별 그룹 배정에서 우수 모델은 고정되지 않았다.

평가단위에 따른 변화가 표현학습 모델에서도 나타나는지 확인하기 위해 소형 CNN을 보조적으로 평가하였다. CNN의 BA는 영상 무작위 분할, 반복영상 제거 후 무작위 분할, 촬영묶음 분할 및 대응 촬영묶음 분할에서 각각 0.524±0.013, 0.517±0.020, 0.484±0.034 및 0.529±0.016이었고, AUROC는 각각 0.533±0.009, 0.510±0.023, 0.475±0.032 및 0.549±0.004였다. CNN에서는 HOG-SVM과 같은 크기의 성능 차이가 나타나지 않았지만, 촬영묶음 분할의 평균 BA와 AUROC가 영상 무작위 분할보다 낮았고 대응묶음 분할에서는 다시 높아졌다. CNN을 포함한 세 모델의 결과는 Fig. 2에 제시하였다.

Fig. 3은 중복 저장영상을 제거하고 촬영묶음당 고유 영상 5장을 모두 사용하였을 때의 촬영묶음별 결정마진 분포를 나타낸다. HOG 모델은 영상 무작위 분할에서 품질등급 0이 음수, 품질등급 1이 양수 영역에 주로 분포하였으나 촬영묶음 분할에서는 등급별 중심의 방향이 반대로 바뀌었다. RBF 모델의 경우 영상 무작위 분할과 촬영묶음 분할 모두에서 반대 방향의 분포를 보였다. 대응 촬영묶음까지 함께 분리하자 두 모델 모두 등급별 중심이 정상 방향으로 이동하였다. 30회 반복에 대한 평균 AUROC는 반복 제거 후 영상 무작위, 반복 제거 후 촬영묶음 및 대응 촬영묶음 분할의 순으로 HOG 모델에서 0.813, 0.246, 0.572였으며 RBF 모델에서는 0.357, 0.215, 0.590이었다. 이 값들은 Table 3의 해당 세 조건과 동일한 데이터에 대한 결과이나, 앞의 두 조건은 분할 반복 횟수가 10회에서 30회로 늘어남에 따라 최대 0.021의 편차를 보인다.

3.4 촬영묶음당 프레임 수의 영향

Fig. 4는 촬영묶음당 고유 프레임 수를 변화시키며 얻은 결과를 나타낸다. 영상 무작위 분할에서 동일 촬영묶음의 노출률은 k =

1에서 0%였으나, k = 2에서 80.5%로 급격히 상승하였고 이후 96.4%, 99.3%, 99.9%로 증가하여 k = 5에서는 거의 모든 테스트 영상이 학습데이터와 촬영묶음을 공유하였다. 이는 촬영묶음당 프레임이 한 장뿐인 경우 무작위 분할과 촬영묶음 분할이 사실상 동일한 조건이 되지만, 프레임이 두 장만 되어도 대부분의 촬영묶음이 학습데이터와 테스트데이터에 걸치게 됨을 의미한다. 한편 대응그룹 노출률은 k = 1에서 이미 80.3%에 달하였고 k = 3부터 100%에 도달하였는데, 이는 등급 간 대응관계가 프레임 반복과 무관하게 촬영묶음 사이에 존재하기 때문이다.

이러한 노출률의 변화는 분류성능에 그대로 반영되었다. HOG 모델의 영상 무작위 BA는 k = 1에서 0.376이었으나 k = 5에서 0.762로 상승한 반면, 동일한 표본에 촬영묶음 분할을 적용한 경우에는 0.409에서 0.288로 오히려 낮아졌다. 그 결과 두 조건의 BA 차이는 k = 1의 −0.033에서 k = 5의 0.475로 확대되었다. 반면 대응 촬영묶음 분할의 BA는 0.542에서 0.559 사이에 머물러 프레임 수의 영향을 거의 받지 않았다.

RBF 모델에서는 영상 무작위 BA가 0.381에서 0.395 사이로 큰 변화를 보이지 않았으나, 촬영묶음 분할 BA는 0.411에서 0.281로 낮아졌고 대응 촬영묶음 분할 BA는 0.515에서 0.579로 완만하게 상승하였다. 즉 프레임 수 증가의 영향은 특징표현과 분할단위에 따라 방향과 크기가 다르게 나타났으며, 두 모델에 공통적으로 관찰된 것은 촬영묶음 분할에서의 성능 저하였다.

4. 고 찰

4.1 중복 저장영상과 독립 평가단위

본 연구에서 중첩 노출률을 산출함으로써 중복 저장영상에 의한 노출과 촬영묶음 공유에 의한 노출을 수치로 구분할 수 있었다. 반복영상을 제거한 뒤 무작위로 분할한 조건에서는 동일 해시를 갖는 영상이 더 이상 남지 않았으나, 모든 테스트영상이 여전히 학습데이터와 촬영묶음을 공유하였으며 이 조건에서 HOG 모델의 BA는 0.771이었다. 반면 동일한 700장을 촬영묶음 단위로 분리하자 두 모델의 BA는 모두 0.3 수준까지 낮아졌다. 따라서 파일 해시를 기준으로 한 중복 제거는 반드시 필요한 정제 절차이기는 하나, 새로운 검사사례에 대한 성능을 평가하기 위한 분할규칙을 대신할 수는 없다.

또한 촬영묶음 분할 전후의 큰 성능 차이는 두 모델 모두에서 관찰되었고 부트스트랩 신뢰구간 역시 0을 포함하지 않았으며, 중복 저장영상을 제거한 촬영묶음 분할에서도 결과가 거의 유지된 점 또한 동일한 해석을 뒷받침한다. 즉 영상 무작위 분할에서 나타난 높은 성능에는 중복 저장영상의 노출과 촬영맥락의 공유가 함께 작용한 것으로 볼 수 있다.

4.2 프레임 수와 평가편향

촬영묶음당 고유 프레임 수를 1장에서 5장까지 늘리는 동안 독립 촬영묶음의 수는 140개로 일정하게 유지되었다. 그럼에도 HOG 모델의 영상 무작위 BA는 동일 촬영묶음 노출률의 상승과 함께 0.376에서 0.762로 높아진 반면, 같은 표본에 촬영묶음 분할을 적용한 BA는 오히려 0.409에서 0.288로 낮아졌다. 이는 동일한 검사사례에 프레임을 추가하는 것이 독립적인 검사사례에 관한 정보를 늘리지 않으며, 분할규칙에 따라 보고되는 성능의 차이만을 확대시킨다는 것을 의미한다. 즉 영상 파일 수의 증가는 데이터의 규모를 키우지만 평가의 신뢰성을 높이지는 않는다.

한편 RBF 모델에서는 영상 무작위 성능이 거의 상승하지 않은 채 촬영묶음 분할 성능만 낮아졌으며, 대응 촬영묶음까지 함께 분리한 조건에서는 두 모델 모두 프레임 수에 따른 변동이 상대적으로 작게 나타났다. 따라서 반복 프레임이 성능에 미치는 영향은 특정한 방향으로 고정되지 않으며, 사용하는 특징표현과 데이터의 그룹구조에 따라 달라진다고 볼 수 있다.

4.3 대응 촬영묶음과 모델 선택

서로 다른 품질등급에 속한 파일기반 대응묶음은 영상내용에서도 높은 유사도를 보였으며, 이 관계를 분할단위에 포함하지 않은 촬영묶음 분할에서는 대응그룹의 68.1%가 학습데이터와 테스트데이터에 걸쳐 있었다. 같은 조건에서 두 SVM의 AUROC가 0.5보다 낮고 평균 결정점수의 방향도 품질등급과 반대로 나타났으며, 대응묶음을 함께 분리하자 AUROC가 0.5를 상회하였다. 즉 점수방향의 변화는 파일구조와 영상 유사도로 확인된 대응묶음의 배정과 함께 관찰되었다. 다만 실제 부품·촬영시점·로트 식별자가 제공되지 않았으므로, 그 물리적 원인까지 특정하지는 않았다.

대응 촬영묶음을 하나의 분할단위로 함께 분리하자 평균 BA와 AUROC는 모두 0.5를 상회하였다. 그러나 두 모델의 평균 성능 차이는 0.016에 불과하였고, 30회의 그룹 배정 가운데 RBF 모델이 우수한 경우가 18회, HOG 모델이 우수한 경우가 10회로 나타나 우수 모델이 배정에 따라 달라졌다. 이는 특정 분할 한 번에서 얻어진 순위만으로 모델을 선정하기 어렵다는 것을 보여준다. 따라서 반복 그룹분할을 수행하고 순위의 빈도와 성능 차이의 분포를 함께 보고하면, 모델 선택이 평가설계에 얼마나 민감한지를 확인할 수 있다.

4.4 반복 촬영자료의 성능보고

기존 외관검사 연구는 검사대상, 영상취득장치, 조명 조건, 처리속도와 정확도 등을 주요 실험조건으로 보고해 왔다[1,2]. 반복 촬영자료의 성능을 해석하려면 이러한 정보와 함께 독립 시험사례의 정의, 반복 구조, 분할단위 및 학습·테스트 중첩 정도가 필요하다.

Table 4 Reporting information for repeated-acquisition inspection data

Item	Required description
Deployment target	New image, acquisition case, workpiece, lot, equipment, or time period
Repeat structure	Images per case and handling of exact or near repeats
Split rule	Grouping key and separation rule between training and test sets
Training-test exposure	Exact-image, acquisition-group, and higher-level group overlap
Evaluation procedure	Prediction unit, aggregation rule, metric, folds, and repetitions

분석에 사용한 최소 보고항목은 Table 4에 정리하였다.

부품 식별자나 취득이력이 기록된 데이터에서는 해당 정보를 그룹키로 사용할 수 있다. 원식별자를 확보할 수 없는 경우에는 파일구조와 영상내용에서 관찰되는 관계를 명시하고, 서로 다른 그룹정의에 대한 민감도를 함께 제시할 수 있다. 전체 영상 수와 독립 시험사례 수, 분할 후 중첩 노출률 및 반복 평가절차를 함께 보고하면 성능 추정치의 적용범위를 판단하기 쉬워진다. 본 연구의 수치는 제공된 비공개 용접영상 세트의 파일기반 구조에서 얻은 결과이며, 물리적 사례 식별자가 확보된 자료에서는 동일한 평가절차로 그 재현 여부를 확인할 수 있다.

4.5 분석범위

본 연구는 단일 비공개 용접검사 데이터에서 관찰된 촬영묶음을 사용하였으며, 파일기반 대응묶음 또한 파일구조와 영상 유사도를 근거로 사후에 구성한 것이다. 따라서 실제 부품, 촬영 주기, 로트 및 설비와의 관계는 원식별자가 확보된 데이터에서 별도로 확인되어야 한다. 또한 보조 비교모델로 사용한 소형 CNN은 사전학습과 하이퍼파라미터 탐색 없이 고정된 조건에서 평가한 것이므로, 더 큰 표현학습 모델에서도 동일한 경향이 나타나는지는 추가 검증이 필요하다. 다른 제조조건에 대한 성능은 적용대상에 맞춘 독립 시험단위를 사전에 정의한 뒤 검증하여야 한다.

5. 결 론

본 연구는 반복 촬영 용접검사 자료에서 중복 저장영상의 보존 여부와 평가단위를 교차하여 분류성능의 변화를 분석하였다. 영상 무작위 분할에서는 중복 저장영상과 촬영묶음 공유가 학습데이터와 테스트데이터 사이에 함께 남았고, 중복 저장 영상을 제거한 뒤에도 촬영묶음 수준의 공유는 100%로 유지되었다. 촬영묶음당 고유 프레임 수를 1장에서 5장으로 늘리는 동안 독립 촬영묶음 수는 140개로 일정했지만, HOG-SVM에서 영상 무작위 분할과 촬영묶음 분할의 BA 차이는 -0.033에서 0.475로 확대되었다. 파일기반 대응묶음을 함께 분리하면 두 SVM의 점수 방향은 회복되었으나, 30회 그룹 배정에서 우수 모델은 고정되지 않았다. 소형 CNN의 BA와 AUROC도 분할단위에 따라 달라졌다. 반복 촬영 자료의 분류성능은 영상 수와 함께 독립 시험사례의 정의, 분할단위 및 중첩 노출률을 제시하여 해석해야 한다.

후 기

이 연구는 한국생산기술연구원 유연생산기술연구개발사업(EH-260002) 및 한국산업기술기획평가원 기계산업핵심기술개발사업(RS-2024-00507879)에 의하여 연구되었음.

References

[1] Ko, K. W., 2025, AI and Machine Vision Integrated System for MLCC Defect Inspection, J. Korean Soc. Manuf. Technol. Eng., 34:4 203-209, https://doi.org/10.7735/ksmte.2025.34.4.203.

[2] Baek, J. W., Jeong, U. C., Kim, B. M., Kim, Y. J., Kim, J., 2025, Dataset Construction and Performance Evaluation of a CNN Based Defect Classifier for Automotive Parts Manufacturing, J. Korean Soc. Manuf. Technol. Eng., 34:6 414-420, https://doi.org/10.7735/ksmte.2025.34.6.414.

[3] Chen, Y., Ding, Y., Zhao, F., Zhang, E., Wu, Z., Shao, L., 2021, Surface Defect Detection Methods for Industrial Products: A Review, Appl. Sci.-Basel, 11:16 7657, https://doi.org/10.3390/app11167657.

[4] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., Crawford, K., 2021, Datasheets for Datasets, Commun. ACM, 64:12 86-92, https://doi.org/10.1145/3458723.

[5] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., Aroyo, L. M., 2021, "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI, Proc. 2021 CHI Conf. Human Factors Comput. Syst., 1-15, https://doi.org/10.1145/3411764.3445518.

[6] Kaufman, S., Rosset, S., Perlich, C., Stitelman, O., 2012, Leakage in Data Mining: Formulation, Detection, and Avoidance, ACM Trans. Knowl. Discov. Data, 6:4 1-21, https://doi.org/10.1145/2382577.2382579.

[7] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J.,

Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., Dormann, C. F., 2017, Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure, Ecography, 40:8 913-929, https://doi.org/10.1111/ecog.02881.

[8] Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S. B., Gotz, M., Hamdan, S., Komeyer, V., Kulkarni, A., Lahnakoski, J., Love, B. C., Raimondo, F., Patil, K. R., 2024, On Leakage in Machine Learning Pipelines, arXiv preprint, arXiv:2311.04179, https://doi.org/10.48550/arXiv.2311.04179.

[9] Ramos, P., Ramos, R., Garcia, N., 2025, Data Leakage in Visual Datasets, Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 6368-6378, https://doi.org/10.1109/ICCVW69036.2025.00661.



Jun-Seok Yun
Researcher at the Manufacturing AI Research Center, Korea Institute of Industrial Technology (KITECH). His research interests include industrial artificial intelligence and computer vision.
E-mail: yunjs@kitech.re.kr



Hong-In Won
Principal Researcher and Chief of the Manufacturing AI Cooperation at the Manufacturing AI Research Center, Korea Institute of Industrial Technology (KITECH). His research interests include manufacturing AI, digital twins, autonomous manufacturing, and trustworthy industrial intelligence, with a focus on cyber-physical systems and AI–digital-twin platforms for manufacturing AI transformation.
E-mail: luvhayym@kitech.re.kr