Beyond the Fold: *Quantifying Split-Level Noise and the Case for Leave One Dataset Out AU Evaluation*

Saurabh Hinduja[1] • Gurmeet Kaur[1] • Maneesh Bilalpur[2] • Jeffrey F. Cohn[2] • Shaun Canavan[3]
[1]CGI Technologies and Solutions Inc [2]University of Pittsburgh [3]University of South Florida

CVPR JUNE 3-7, 2026 DENVER COLORADO

Scan for paper

Summary

Subject-exclusive cross-validation *itself* produces measurable performance variance. On BP4D+, the empirical 95% noise floor is ±**0.065 average F1** — larger than most reported SOTA gains. We propose Leave One Dataset Out (LODO) evaluation with subject-level bootstrap as a more stable, interpretable protocol.

The Problem

- Reported AU-detection gains are often **+0.01 to +0.02 F1**.
- Yet the *evaluation protocol* can swing F1 by more than that.
- Same model, same data, same code — different fold ⇒ different metric, sometimes different ranking.
- Single-dataset CV does not capture domain shift either.

Contributions

1. Treat CV performance as a random variable; quantify per-AU 95% noise margins ($\pm 1.96\sigma$).
2. Show F1 has **higher partition sensitivity** than AUC; rankings can flip across splits.
3. Introduce **LODO + subject-level bootstrap** (1000 iterations) for cross-dataset robustness with statistical confidence.

Setup

Studies: split-level variance (BP4D+) and cross-dataset robustness via Leave One Dataset Out (LODO).
Datasets: split-level: BP4D+; LODO: BP4D/BP4D+/DISFA/GFT/UNBC.
Split-level backbones: ResNet50, MobileViT, VGG16.
Protocol: 3-fold subject-exclusive; 4 random partitions; fixed threshold 0.5.

Definitions

Per-AU instability margin:

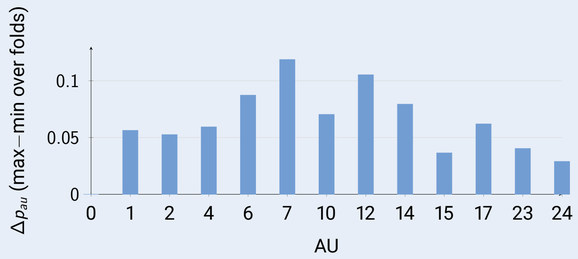
$$\text{Noise}_{au} = \pm 1.96\,\sigma_{au}, \quad \sigma_{au} = \text{StdDev}\left(m_{au}^{(1)}, \ldots, m_{au}^{(K)}\right)$$

Noise floor: mean of per-AU 95% margins across AUs.

Domain Sensitivity (DS): % of LODO transfers in which Δ is significant ($p < 0.05$) under 1000-iteration inter-subject bootstrap.

Split-level Study (BP4D+): Subject Reassignment Distorts the Test Distribution

Identity-disjoint folds preserve subject independence but *not* class distribution. AU prevalence shifts substantially from fold to fold; F1 inherits the shift directly.



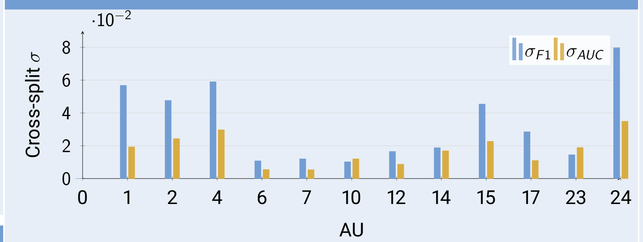
AU7 and AU12 swing by > 0.10 across just three folds.

Split-level Study (BP4D+): Per-AU Noise Floor (ResNet50, 3-Fold)

AU	F1 mean	σ_{F1}	F1 $\pm 1.96\sigma$	AUC mean	AUC $\pm 1.96\sigma$
1	0.454	0.057	±0.111	0.831	±0.038
2	0.366	0.048	±0.093	0.793	±0.048
4	0.329	0.059	±0.116	0.812	±0.058
6	0.860	0.011	±0.021	0.936	±0.011
7	0.886	0.012	±0.024	0.909	±0.011
10	0.904	0.010	±0.020	0.937	±0.023
12	0.894	0.017	±0.032	0.943	±0.017
14	0.820	0.019	±0.037	0.807	±0.033
15	0.436	0.045	±0.089	0.841	±0.045
17	0.502	0.029	±0.056	0.847	±0.022
23	0.562	0.015	±0.028	0.853	±0.037
24	0.213	0.080	±0.156	0.874	±0.068

Average 95% margin: ±0.065 F1 | ±0.034 AUC.

Split-level Study (BP4D+): F1 is ∼2× More Volatile Than AUC



Volatility ratio $\rho = \sigma_{F1}/\sigma_{AUC}$ exceeds 2.0 for AU 1, 4, 7, 15, 17, 24. Threshold-free metrics (AUC) absorb prevalence shifts; threshold-tied metrics (F1) amplify them.

Smaller Folds ⇒ Larger Variance ($\text{Var} \propto 1/n$)

	3-fold	5-fold	Change
Mean σ_{F1}	0.0333	0.0506	↑52%
F1 95% margin	±0.065	±0.099	wider
Mean σ_{AUC}	0.0174	0.0261	↑50%
AUC 95% margin	±0.034	±0.051	wider

Going from 33% to 20% test coverage roughly halves n per fold; per-fold estimation variance — and thus cross-fold σ — rises accordingly. The relative ordering F1 ≫ AUC is preserved.

Recent BP4D+ SOTA — All Within the Noise Band

Model	Year	F1
FMAE-IAT	2024	0.668
VisAULa	2026	0.667
MDHR	2024	0.666
AUFormer	2024	0.662
MCM	2024	0.660
CLEF	2023	0.659
ANFL	2022	0.655
STIF-DA	2025	0.654
AQ-CSL	2024	0.649
KSRL	2022	0.645
PIAP	2021	0.644
JAA-Net	2021	0.627

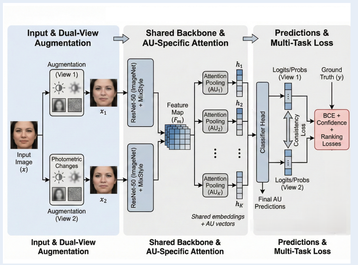
Best−worst spread = 0.041 F1; best−median = 0.019 F1. **Both fall below the ±0.065 noise floor** — much of the apparent ranking is not reliably distinguishable under repeated splits.

Implication

Any F1 improvement smaller than ≈0.06 average F1 on BP4D+ *cannot* be reliably distinguished from partition-induced variability under the same protocol.

Leave One Dataset Out (LODO)

Train on $\cup_{i \neq k} D_i$; test on held-out D_k. *No* target-domain frames, statistics, or thresholds. Repeat for each of 5 datasets.



LODO F1 Across Held-Out Datasets

Held-out	AU6	AU7	AU10	AU12	AU14	AU24
BP4D	.797	.803	.847	.863	.686	.238
BP4D+	.848	.852	.885	.876	.669	.328
DISFA	.588	—	—	.760	—	—
GFT	.742	.556	.527	.769	.084	.146
UNBC	.419	.224	.036	.485	—	—

High-prevalence AUs (6/7/10/12) transfer best; low-prevalence AUs and absent annotations destabilize F1. AUC remains noticeably more stable than F1 across transfers.

Domain Sensitivity (Subject-Level Bootstrap, N=1000)

AU	ΔF1	ΔAUC	F1 DS	AUC DS
1	−0.172	−0.065	67%	33%
6	−0.172	−0.028	**100%**	60%
7	−0.386	−0.027	**100%**	67%
10	−0.472	−0.071	**100%**	50%
12	−0.167	−0.044	80%	40%
14	−0.307	−0.119	67%	67%
17	−0.147	−0.042	**100%**	0%

Operating-point F1 is disrupted in nearly every transfer; AUC is far more stable. The optimal decision threshold is *domain-dependent*.

Recommendations for AU Evaluation

1. Report variance and confidence intervals across repeated splits.

2. To measure robustness to domain shift, add a multi-dataset protocol (LODO).

3. Report *both* threshold-free (AUC) and operating-point (F1) metrics.

4. Treat sub-noise gains as statistically indistinguishable, not SOTA.

Contact: saurabhh@ntweat.com • **CVPR 2026 Workshop | Denver, Colorado | June 3–7, 2026**