Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Singleton-Set Gating versus Confidence Thresholding in RGB–IMU Activity Recognition

David Shulman

dshulman@campus.haifa.ac.il

University of Haifa

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11750091/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Singleton-Set Gating versus Confidence Thresholding in RGB–IMU Activity Recognition

David Shulman[1*]

[1*]Department of Statistics, University of Haifa, Haifa, Israel.

Corresponding author(s). E-mail(s): david.shulman.research@gmail.com;

Abstract

Accepting a classifier prediction only when a confidence detector passes and a calibrated prediction set contains one label is an intuitive rejection rule. Its performance depends on both decision geometry and calibration. We examine this rule in RGB–IMU activity recognition on the University of Texas at Dallas Multimodal Human Action Dataset (UTD-MHAD), with eight people, five held-out-action episodes and three person schedules. Two video encoders share a sensor branch and use arithmetic or normalized-product fusion under one training seed. We compare singleton-gated and confidence-only acceptance at the same accepted-correct calibration count and threshold endpoint. At each deployment's original composite count, confidence-only acceptance reduces arithmetic-fusion unknown acceptance by 8.93 and 10.61 percentage points for the recurrent and Transformer systems, reaching 26.16% and 27.40%. Mean correct-known changes are +0.91 and +0.47 points, but 40 and 47 of 120 deployments lose useful output. A paired fixed-target analysis shows smaller, operating-point-dependent differences: gains disappear at low targets and composite feasibility declines as the target increases. Product-fusion rule differences are small; its original operating points have substantially greater known and unknown acceptance than arithmetic fusion. Threshold geometry and conditional exclusions distinguish filtering at a fixed cutoff from calibration-induced cutoff movement. The findings support matched-calibration, support-aware evaluation of singleton gating rather than a uniform preference across operating points.

Keywords: activity recognition; selective classification; prediction sets; multimodal fusion; calibration transfer

1 Introduction

An RGB–IMU activity recognizer with an abstention option must separate reliable known-action predictions from within-vocabulary errors and out-of-vocabulary actions. In set-valued classification, a prediction set represents class ambiguity by returning multiple candidate labels, whereas a singleton set isolates a single candidate. Requiring a prediction to pass a top-one confidence detector and simultaneously form a singleton prediction set therefore offers a natural composite rule for withholding ambiguous decisions. Prediction-set validity, however, targets marginal coverage of the true known label, whereas selective acceptance is evaluated by accepted-correct yield, selective risk and unknown rejection.

When a singleton filter is applied at a fixed confidence cutoff, it can exclude both unknown actions and correctly classified known actions. Once the confidence cutoff is recalibrated on known calibration data to restore the required accepted-correct count, the composite

rule operates at a lower top-one threshold than a confidence-only rule calibrated to the same target. Thus the net held-person effect of singleton gating depends on the interaction between simplex decision geometry and threshold recalibration rather than on fixed-cutoff filtering alone.

To isolate this interaction, we evaluate an initial detector-plus-singleton composite policy, a recalibrated composite policy and a recalibrated confidence-only policy on identical fitted branch predictions. Matching the accepted-correct calibration count and endpoint convention decomposes policy replacement into a recalibration component and a matched-calibration rule contrast. Reporting absolute outcome rates, deployment losses, conditional exclusions and paired fixed-target sensitivity characterizes both where the rules differ and how attainable support changes across operating points.

2 Related work

Set-valued and selective classification optimize different objectives. Sadinle, Lei and Wasserman [1] characterize label-probability level sets minimizing expected size under coverage constraints. Romano et al. [2] study adaptive prediction sets. Singleton-oriented uncertainty triage also appears in pathology: Olsson et al. [3] measure correct-single-label efficiency and flag ambiguous or atypical biopsies for human intervention. This motivates examining selective use of sets, while differing from the present detector-plus-singleton HAR rule. Geifman and El-Yaniv [4] study risk–coverage guarantees; our calibration targets an empirical accepted-correct count.

Xia and Bouganis [5, 6] evaluate selective classification jointly with out-of-distribution rejection. Their SIRC method combines softmax and feature information, and the journal extension considers multiple secondary scores. Their dependence on unknown prevalence and error costs motivates reporting component rates here. Jaeger et al. [7] emphasize classifier failures and distribution shifts in confidence evaluation. These studies motivate the evaluation objective rather than establish this particular composite rule as a standard baseline.

Dhamija et al. [8] introduced open-set classification-rate curves. Vaze et al. [9] show that strong closed-set classifiers can yield effective open-set baselines; maximum softmax probability is a standard comparator [10]. Our operational thresholds use known calibration people, while test-swept curves are descriptive. Open-set wearable settings include hierarchical attention [11] and head-mounted inertial sensing [12].

Classical classifier combination [13] and RGB–IMU recognition [14, 15] predate this study. MultiOOD [16] trains with modality discrepancy and synthetic outliers; our negative-Jensen–Shannon baseline reuses unchanged probabilities. Recent modality-aware approaches include continual egocentric novelty detection with replay-based reliability [17] and a post-hoc detector using unimodal/fused discrepancy and features [18]. Those methods use different training or signals. We isolate the interaction between a singleton gate and known-only calibration in a compact, fixed RGB–IMU system.

3 Data and classifiers

3.1 Dataset and person-disjoint design

UTD-MHAD [19] contains recordings from eight people and 27 actions; the paired inventory used here has 861 clips, with three expected recordings absent. Five fixed class episodes each designate 20 actions as known and seven as unknown. Within each of three person schedules, four people train the branches, three different people calibrate acceptance, and the remaining person is tested. Rotating the test person yields 120 deployments per encoder/fusion system. All open-set models use seed 17. The two encoders share the fitted sensor branch, so their results are neither independent datasets nor independent sensor replications.

Unknown classes are excluded from model training, normalization and threshold fitting. Each test person's actions are evaluated under the same branch predictions across compared policies. Repeated schedules and class episodes reuse people and recordings, so 120 deployments do not represent 120 independent experimental units. Each deployment tests 79 or 80 known clips and 27 or 28 unknown clips: 30 deployments have 79/28, 15 have 80/27 and 75 have 80/28. Missing recordings and literal class/person assignments are listed in Online Resource 1.

The single wearable is mounted on the wrist for actions 1–21 and thigh for actions 22–27. Placement and activity group are therefore confounded. We report strata separately and repeat threshold calibration and evaluation using only wrist actions. These wrist-only analyses retain branches trained on the original mixed-placement data; they are not a retraining experiment or a causal estimate of mounting location.

3.2 Representations and fitted networks

Each inertial clip retains the native six-channel order and is interpolated on normalized sample index to 128 steps, with training-only channel normalization. Video is resized to 160×120 pixels; dense Farnebäck flow [20] is summarized as a $[16, 128]$ histogram-of-optical-flow (HOF) tensor: 16 temporal bins, a 4×4 spatial grid and eight angular bins. No depth, skeleton, privileged teacher or pretrained weights are used.

The shared sensor branch has two temporal convolutions and a pooled classifier. The recurrent video branch uses a single 64-unit gated recurrent unit (GRU) [21]; the Transformer [22] uses width 64, one four-head encoder layer and temporal averaging. Full layer specifications, flow parameters and normalization rules are retained in Online Resource 1, Section S2.

The $K = 20$ sensor, GRU and Transformer heads contain 29,428, 38,804 and 40,308 parameters, yielding fused totals of 68,232/69,736. AdamW uses learning rate and weight decay 0.001 with unweighted cross-entropy; sensor/video training uses 60/40 epochs and batches of 32/16. Fusion adds no parameters. Full initialization, backend and optimization specifications are in Online Resource 1.

3.3 Probability fusion

Let s_k and v_k be sensor and video probabilities for the same ordered class k. Arithmetic fusion is $p_A(k) = (s_k + v_k)/2$. Product fusion is the normalized distribution

$$p_P(k) = \frac{\exp(a_k - \max_j a_j)}{\sum_j \exp(a_j - \max_l a_l)},$$
$$a_k = \log \max(s_k, \epsilon) + \log \max(v_k, \epsilon),$$

where $\epsilon = 10^{-12}$. Both rules use the same fitted logits and label order. Their effects are evaluated empirically; no independence of branch errors is assumed.

4 Acceptance and calibration

4.1 Initial composite and its geometry

Write the ordered fused probabilities as $p_1 \geq p_2 \geq \cdots \geq p_K$. The initial composite accepts when the confidence detector passes, $1 - p_1 \leq \tau$, and the set $\Gamma = \{k : 1 - p_k \leq q\}$ contains exactly one label. Known-calibration scores $1 - p_1$ and $1 - p_{\text{true}}$ determine τ and q at order-statistic ranks $\lceil 0.95(n+1) \rceil$ and $\lceil 0.90(n+1) \rceil$, respectively, where n is the pooled known-calibration count. A rank above n gives an infinite threshold.

For finite thresholds the exact real-arithmetic acceptance region is

$$p_1 \geq \max(1 - \tau, 1 - q), \qquad p_2 < 1 - q. \tag{1}$$

Membership is inclusive, so equality in the second condition includes a second label. Infinite q includes every label and rejects all examples when $K \geq 2$; infinite τ removes only the detector restriction. Numerical replay preserves the original subtraction predicates at floating-point boundaries.

Since $p_2 \leq 1 - p_1$, either $q < 0.5$ or $\tau + q < 1$ suffices to make the second condition redundant after the combined top-one lower bound. The inequalities are strict: $p_1 = p_2 = 0.5$ can pass a non-strict top-one bound but fail singleton membership. We distinguish redundancy over the probability simplex, absence of observed test exclusions at fixed thresholds, and equality after separate recalibration. The supplement gives boundary examples and the exact redundancy criterion.

4.2 Matching calibration counts and separating comparisons

Let r be the exact number of correctly classified known calibration examples accepted by the initial composite. Three procedures are evaluated on identical test rows:

1. **Initial composite:** the detector and set thresholds defined above.
2. **Recalibrated composite:** the same fixed singleton eligibility rule, with its confidence cutoff fitted to attain r correct calibration acceptances.
3. **Recalibrated confidence:** confidence alone, fitted to that same r with the same endpoint convention.

For each rule's eligible correctly classified calibration examples, the primary convention selects the r-th largest confidence score and accepts all eligible examples with confidence at least as large, defining the strictest inclusive threshold attaining at least r correct acceptances. Tied boundary scores are accepted together, with tie overshoot recorded; a requested count above the eligible correct count is marked infeasible, $r = 0$ yields reject-all and the maximum attainable count selects the minimum eligible score. Throughout these comparisons, the original set threshold q is held fixed rather than jointly tuned.

For each metric M, the replacement difference between recalibrated confidence (C) and the initial composite (I) decomposes additively within every deployment via the recalibrated composite (R):

$$M(C) - M(I) = [M(R) - M(I)] + [M(C) - M(R)], \quad (2)$$

where $M(R) - M(I)$ captures recalibration under a fixed singleton gate and $M(C) - M(R)$ isolates the matched-calibration rule contrast. Sensitivity analyses match total accepted-known calibration counts or place cutoffs at correct-count plateau midpoints (Online Resource 1, Section S3). All threshold fitting uses known calibration examples exclusively.

4.3 Outcomes, support and uncertainty

Correct-known acceptance (CCR) and accepted-wrong rate divide accepted correct and accepted incorrect known queries by all known test queries; their sum equals known coverage. Known selective risk divides accepted-wrong queries by all accepted-known queries and is undefined at zero coverage. Unknown false acceptance divides accepted unknown queries by all unknown test queries, whereas prediction-set coverage is the fraction of known true labels in Γ. Because error costs and unknown prevalence vary by application, we report all five component outcomes rather than a single weighted risk.

To respect the person-disjoint structure, outcomes and contrasts are averaged first within each held-out test person and then equally across the eight people. Pairwise rule comparisons restrict to the intersection of feasible deployment identities before averaging, and an eight-person primary estimate requires at least one feasible paired deployment per person (with finite risk under both rules for selective-risk contrasts).

Descriptive 95% person-bootstrap intervals use 5,000 resamples of people that preserve all repeated deployments within each selected person, complemented by seven-degree-of-freedom paired t intervals across person means, leave-one-person-out ranges and crossed person-by-episode resampling. Given the eight-person cohort, overlapping schedules, fixed episodes and retrospective exploration, these intervals summarize variability rather than confirmatory population guarantees: an interval spanning zero does not establish equivalence, and a favorable mean can coexist with individual deployment losses.

4.4 Empirical regime checks

To explain when the singleton constraint is active, we retain all 120 deployments per system and record fitted thresholds (τ, q), branch top-1 agreement, fused top-two probabilities (p_1, p_2) and conditional singleton exclusion rates among initial detector passes for known-correct, known-wrong and unknown queries across wrist and thigh strata, evaluating each test query against its own deployment's thresholds.

5 Results

5.1 Matched-calibration rule comparison

Table 1 gives absolute levels for all three procedures; Table 2 compares recalibrated confidence against the recalibrated composite, using all 120 paired deployments and eight people per system. For arithmetic fusion, unknown acceptance

Table 1 Absolute test outcome levels (%). I: initial composite; R: recalibrated composite; C: recalibrated confidence. Within each system, R and C match that deployment's initial accepted-correct calibration count. Means first average deployments within person, then equally across eight people (120 deployments). Selective risk is averaged as a deployment-level ratio; it is not the ratio of the displayed mean rates.

System	Policy	Correct-known	Accepted-wrong	Coverage	Selective risk	Unknown
GRU / arithmetic	I	65.26	4.36	69.62	5.84	37.03
GRU / arithmetic	R	64.92	3.94	68.87	5.31	35.09
GRU / arithmetic	C	65.83	2.80	68.62	3.74	26.16
GRU / product	I	83.65	7.17	90.82	7.83	72.01
GRU / product	R	83.28	6.94	90.22	7.62	70.49
GRU / product	C	83.31	6.94	90.25	7.62	70.79
Transformer / arithmetic	I	70.93	5.01	75.94	6.29	40.49
Transformer / arithmetic	R	70.63	4.52	75.15	5.66	38.01
Transformer / arithmetic	C	71.10	3.00	74.10	3.82	27.40
Transformer / product	I	85.84	6.51	92.36	7.03	73.60
Transformer / product	R	85.56	6.29	91.85	6.83	72.16
Transformer / product	C	85.60	6.30	91.91	6.84	72.13

decreases by 8.93 percentage points with the GRU and 10.61 points with the Transformer. Accepted-wrong known predictions decrease by 1.15 and 1.52 points, and known selective risk decreases by 1.57 and 1.84 points. Mean correct-known acceptance increases by 0.91 and 0.47 points. Thus the average rejection gains are not explained by an average reduction in accepted-correct output in this particular comparison.

The useful-output result requires caution. Paired t intervals for the arithmetic correct-known differences are [−0.01, 1.82] and [−0.19, 1.13] points, both including zero; no noninferiority margin was specified. Confidence-only acceptance loses accepted-correct output in 40/120 GRU and 47/120 Transformer deployments, with worst changes −10.00 and −10.13 points. Unknown acceptance increases in 14 GRU and 18 Transformer deployments; its GRU change ranges from −42.86 to +39.29 points. All eight person means are negative (Online Resource 1, Fig. S1).

At its own initial-count operating point, product fusion has substantially greater useful output and unknown acceptance: roughly 83–86% CCR and 70–74% unknown acceptance (Table 1). The matched-rule unknown-acceptance change is +0.30 points for the GRU and −0.03 for the Transformer, with very small changes in known outcomes. The GRU percentile interval excludes zero, but its paired t interval [−0.07, 0.68] does not. These small, endpoint-sensitive rule effects must be distinguished from comparisons between fusion rules at a common target.

5.2 Replacement versus recalibration

Figure 1 separates the three contrasts. Replacing the initial composite with recalibrated confidence reduces arithmetic unknown acceptance by 10.87 and 13.10 points. Recalibrating the composite itself accounts for 1.94 and 2.49 points; the remaining 8.93 and 10.61 points are the matched-rule differences. For product fusion, recalibration reduces unknown acceptance by 1.52 and 1.44 points, whereas the matched-rule differences are +0.30 and −0.03. The practical replacement reductions of 1.22 and 1.46 points therefore mostly reflect recalibration for product fusion. Calling the full replacement effect a singleton-gate effect would misidentify the comparison.

5.3 Calibration and deployment sensitivity

Matching total accepted-known calibration count gives arithmetic unknown-acceptance reductions of 9.08 and 10.83 points; matching correct counts at plateau midpoints gives reductions of 11.47 and 13.76 points. The direction therefore persists under these two alternative conventions, while its magnitude changes. Product midpoint differences become −0.59 and −0.72 points, reversing the GRU strict-endpoint direction; both paired t intervals include zero. The comparison is not invariant to an admissible cutoff position.

Wrist-only calibration and evaluation, with the originally mixed-placement-trained branches,

Table 2 Matched-calibration rule differences, confidence minus composite. Values are percentage-point changes in equal-person means, with descriptive 95% person-bootstrap intervals. Correct-count calibration and the strictest inclusive endpoint are used. All entries retain 120 deployments and eight people; risks are defined for every paired cell. Negative unknown acceptance, accepted-wrong and risk differences are favorable; a lower coverage value alone is not an improvement.

System	Correct-known	Accepted-wrong	Coverage	Selective risk	Unknown
GRU / arithmetic	+0.91 [+0.24, +1.65]	−1.15 [−1.72, −0.68]	−0.24 [−0.87, +0.33]	−1.57 [−2.35, −0.93]	−8.93 [−12.21, −5.58]
GRU / product	+0.03 [−0.04, +0.10]	0.00 [−0.03, +0.03]	+0.03 [−0.04, +0.12]	0.00 [−0.04, +0.03]	+0.30 [+0.06, +0.62]
Transformer / arithmetic	+0.47 [−0.01, +1.03]	−1.52 [−2.15, −0.94]	−1.05 [−1.91, −0.18]	−1.84 [−2.63, −1.12]	−10.61 [−13.03, −8.01]
Transformer / product	+0.04 [−0.02, +0.11]	+0.01 [0.00, +0.03]	+0.05 [−0.01, +0.12]	+0.01 [0.00, +0.03]	−0.03 [−0.09, 0.00]

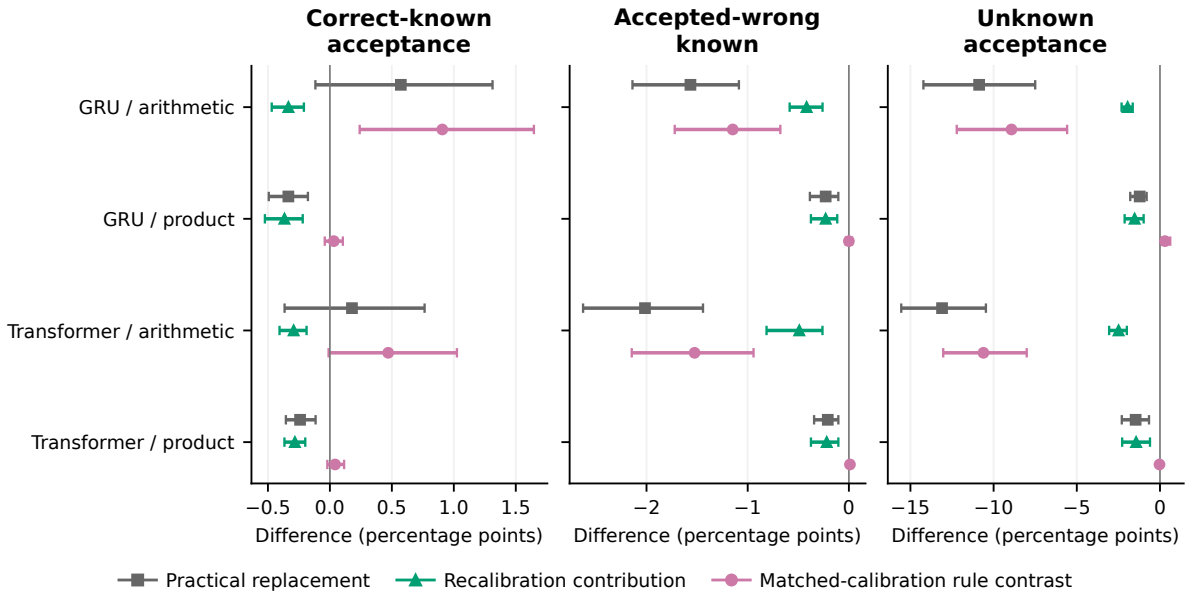


Fig. 1 The practical replacement difference equals the recalibration contribution plus the matched-calibration rule difference within each deployment. Points are equal-person means; bars are descriptive 95% person-bootstrap intervals from the same 120 deployments and eight people per system. Intervals do not add arithmetically. Correct-known acceptance, accepted-wrong and unknown acceptance use their denominators from Section 4.3; Table 2 retains coverage and selective risk.

gives arithmetic unknown-acceptance reductions of 5.38 [2.92, 8.13] and 8.68 [6.75, 10.13] points. Correct-known changes are +0.40 [−0.18, 0.96] and +1.23 [0.22, 2.24]. This reduces the concern that the entire finding is a thigh-action artifact, but does not isolate sensor placement from activity or retrain away the confounding.

The GRU arithmetic unknown reduction is only 0.45 points in episode U3, versus 8.66–13.06 in the other episodes. Transformer episode reductions range from 7.78 to 12.85 points. The supplement retains dependence sensitivities and person-level results.

5.4 Threshold regimes and exclusions

Figure 2 shows every initial threshold pair. The sufficient redundancy union holds in 56/120 and 75/120 product deployments, and in none of the arithmetic deployments. On held-out examples the composite equals its combined top-probability

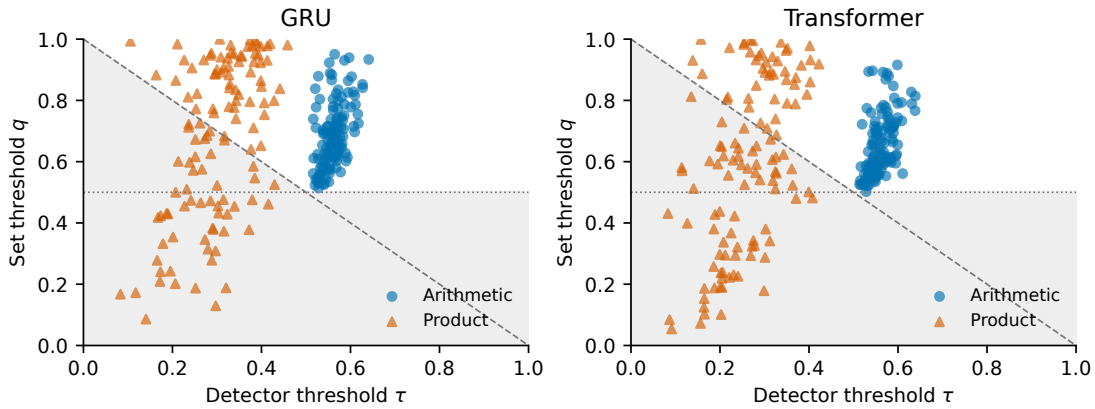


Fig. 2 All 120 initial threshold pairs per system. Shading marks the open sufficient region $q < 0.5$ or $\tau + q < 1$ where the second-probability restriction is redundant relative to the combined top-one cutoff. Boundary lines are excluded from the sufficient condition. The detector and top-inclusion bounds can still differ. Conditional exclusion rates at these same thresholds are reported in Online Resource 1, Figure S4.

cutoff in 62 and 82 product deployments. Separately recalibrated composite and confidence policies produce identical complete test acceptance masks in 107 and 111 product deployments, compared with two and one arithmetic deployments. The three counts address algebraic redundancy, observed inactivity and calibrated-policy equality, respectively; they are not interchangeable.

At the original detector cutoff, the arithmetic singleton filter excludes 21.82%/16.27% of correct-known detector passes and 54.67%/50.09% of unknown detector passes (GRU/Transformer), using person-first conditional rates. The underlying repeated-query counts are 1,740/7,985 and 1,323/8,110 correct-known exclusions, and 1,530/2,769 and 1,373/2,728 unknown exclusions; pooled ratios differ from person-first means. These held-person exclusions quantify fixed-cutoff filtering. On calibration rows, imposing the gate at the confidence-only cutoff loses correct acceptances in 120/119 cells. Restoring the common count lowers the composite cutoff: the confidence-only cutoff is higher in 120/120 and 119/120 arithmetic deployments, with median paired gaps 0.081 and 0.063. Its more restrictive top-one cutoff can reject more held-person unknowns despite dispensing with the singleton filter. Final outcomes and deployment losses, rather than fixed-cutoff exclusion rates alone, determine that comparison.

Unknown branch top-1 agreement is 18.18%/14.32% (GRU/Transformer). Fused top-two means change from $(0.548, 0.308)/(0.532, 0.317)$ under arithmetic fusion to $(0.881, 0.095)/(0.885, 0.092)$ under product fusion. This measured concentration helps explain product-gate inactivity; the supplement gives a counterexample to universal sharpening.

5.5 Fixed-target sensitivity

To test whether the matched-rule contrast persists away from each deployment's adaptive initial-composite count, we recalibrate both rules across fixed calibration CCR targets (0.50–0.85), keeping each deployment's original set threshold q fixed and applying the strictest inclusive endpoint on identical paired support. Required accepted-correct counts are exact decimal ceilings of the target times the calibration-known count (238–240 queries), giving zero tie overshoot across all $5,566$ feasible cells out of $6,720$ evaluations ($1,154$ infeasible; Online Resource 1, Section S10, Tables S11–S13).

Table 3 shows that the arithmetic rule contrast depends strongly on the operating point: paired unknown-acceptance differences are near zero at 0.50–0.60 and reach $-3.72/-1.14$ points at 0.65 and $-3.58/-3.29$ at 0.70, while composite feasibility drops from 105/8 and 113/8 at 0.50 to 4/4 and 10/5 at 0.85. Because no fixed target retains all 120 deployments, the 8.93- and 10.61-point reductions at the adaptive count are not uniform fixed-target effects. Holding the 0.85-feasible subsets (4/4, 10/5) fixed at lower targets yields zero

Table 3 Paired arithmetic rule differences ($C - R$, percentage points) and common support (N deployments / P people) across fixed calibration CCR targets ($P < 8$: conditional subset means).

	GRU / arithmetic			Transf. / arithmetic		
Target	ΔCCR	ΔUnk.	N/P	ΔCCR	ΔUnk.	N/P
0.50	0.00	+0.06	105/8	0.00	0.00	113/8
0.60	+0.15	−0.95	85/8	+0.09	−0.13	100/8
0.65	+0.29	−3.72	71/8	+0.04	−1.14	89/8
0.70	+1.57	−3.58	48/8	+0.55	−3.29	73/8
0.75	+0.25	−4.35	33/7	+0.38	−4.41	53/8
0.80	+0.01	−4.16	19/5	+0.62	−3.88	27/6
0.85	+2.81	−4.46	4/4	+0.56	−6.02	10/5

unknown differences through 0.80 (GRU) and 0.75 (Transformer; +1.33 at 0.80).

For product fusion, paired rule differences are 0.00 across all outcomes through 0.80 (89/8 to 120/8) and +0.36/0.00 for unknown acceptance at 0.85 (67/8, 83/8). Comparing fusion rules at a common confidence-only target retains all 120/8 deployments through 0.75: arithmetic versus product unknown acceptance is 24.02% versus 25.16% (GRU) and 19.18% versus 22.01% (Transformer) at 0.65, and 29.90% versus 32.40% and 24.36% versus 28.47% at 0.70, confirming that the $\sim 72\%$ product unknown acceptance in Table 1 reflects its permissive initial operating point (83–86% CCR).

5.6 Recognition and cost controls

In a 27-action, three-seed leave-one-person-out closed-set control, Transformer product fusion reaches 91.68% mean accuracy (85.92%/89.17% for GRU arithmetic/product and 87.85% for Transformer arithmetic), contextualizing our lightweight branches against published UTD-MHAD results of 95.6% [14] and 92.33% [15] under different architectures and protocols. Replacing the initial composite with a negative branch Jensen–Shannon divergence score at matched calibration count reduces arithmetic unknown acceptance by 9.29 and 10.25 points (+0.38/ + 0.04 CCR) and product unknown acceptance by 6.07 and 1.66 points (−0.50/ − 0.48 CCR), whereas training-standardized energy increases unknown acceptance by 8.84–15.43 points (Online Resource 1, Table S6). Across 360 rotating one-person temperature / two-person threshold allocations per system, post-fusion temperature scaling fits median temperatures $T = 0.871/0.782$ (arithmetic) and $T = 2.500/2.482$ (product), altering confidence rankings (mean Kendall $\tau \in [0.886, 0.918]$) and reducing product unknown acceptance by 5.02 and 5.31 points ($-1.39/-1.28$ CCR) while increasing arithmetic unknown acceptance by 3.07 and 4.35 points ($+1.01/+1.51$ CCR; Section S6, Table S8). Single-thread CPU latency is 433.5 ms (GRU) and 427.1 ms (Transformer) per clip, dominated by video decoding and flow.

6 Discussion

The interaction between fixed-cutoff filtering and calibration explains the composite rule's behavior. At the initial detector threshold, adding the singleton condition removes roughly half of remaining unknown passes under arithmetic fusion, but also removes 16–22% of correct-known passes. When both policies are calibrated to the same accepted-correct count, compensating for those excluded known predictions forces the composite policy to adopt a lower top-one cutoff than confidence thresholding alone (0.081 and 0.063 median gap), outweighing the singleton filter on held-out test people.

At the same time, the paired fixed-target analysis shows that this arithmetic advantage is concentrated at each deployment's adaptive initial-composite count, where the initial set threshold q admits all 120 deployments. Under fixed calibration targets, the paired rule difference shrinks at moderate targets, vanishes at low targets and is accompanied by declining composite feasibility near the singleton-eligibility ceiling. Under product fusion, strong probability concentration leaves the singleton condition mostly inactive, and common-target comparisons show that its high initial unknown acceptance is primarily an operating-point effect. Composite rules combining set-size conditions with confidence detectors should therefore be evaluated with their calibration procedure by reporting absolute rates, matching calibration counts and endpoints, and tracking deployment losses and infeasible targets.

7 Limitations

The evidence comes from one eight-person, segmented-clip benchmark and one open-set seed;

both video encoders share a sensor branch. Person-disjoint fitting prevents direct use of the test person but leaves dependence across deployments and retrospective benchmark exposure. Unknowns are withheld classes within the same acquisition setting. Placement and action group are confounded, and wrist-only calibration retains mixed-placement training. Additional seeds would test training variability, while new participants and settings are needed for transfer. These conditional results establish neither uniform rule superiority nor a failure of prediction-set coverage.

8 Conclusion

For these fixed RGB–IMU classifiers, a matched-calibration comparison separates singleton-rule effects from recalibration. Confidence-only acceptance reduces arithmetic unknown acceptance at the original reference counts, with smaller, support-dependent effects over fixed targets and useful-output losses in individual deployments. Product differences are small at matched targets, making absolute rates and attainable support essential to interpretation.

Supplementary information

Online Resource 1 provides detailed methods, sensitivity analyses, temperature controls and supplementary figures and tables.

Statements and Declarations

Author contributions

David Shulman: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing.

Acknowledgements

Computing resources were provided by the Israel Data Science and AI Initiative (IDSAI) and the Israeli Council for Higher Education (CHE), with infrastructure managed by the Inter-University Computation Center (IUCC).

AI assistance

Generative AI and coding tools assisted provenance auditing, implementation, figures/tables and manuscript drafting. The author is responsible for verification of all code, results and text.

References

[1] Sadinle, M., Lei, J., Wasserman, L.: Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association **114**(525), 223–234 (2019) https://doi.org/10.1080/01621459.2017.1395341

[2] Romano, Y., Sesia, M., Candès, E.: Classification with valid and adaptive coverage. In: Advances in Neural Information Processing Systems, vol. 33 (2020)

[3] Olsson, H., Kartasalo, K., Mulliqi, N., Capuccini, M., Ruusuvuori, P., Samaratunga, H., Delahunt, B., Lindskog, C., Janssen, E.A.M., Blilie, A., ISUP Prostate Imagebase Expert Panel, Egevad, L., Spjuth, O., Eklund, M.: Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. Nature Communications **13**, 7761 (2022) https://doi.org/10.1038/s41467-022-34945-8

[4] Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

[5] Xia, G., Bouganis, C.-S.: Augmenting softmax information for selective classification with out-of-distribution data. In: Asian Conference on Computer Vision (2022)

[6] Xia, G., Bouganis, C.-S.: Augmenting the softmax with additional confidence scores for improved selective classification with out-of-distribution data. International Journal of Computer Vision **132**, 3714–3752 (2024) https://doi.org/10.1007/s11263-024-02029-3

[7] Jaeger, P.F., Lüth, C.T., Klein, L., Bungert, T.J.: A call to reflect on evaluation practices for failure detection in image classification. In: International Conference on Learning Representations (2023). https://arxiv.org/abs/2211.15259

[8] Dhamija, A.R., Günther, M., Boult, T.E.: Reducing network agnostophobia. In: Advances in Neural Information Processing Systems, vol. 31 (2018)

[9] Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: A good closed-set classifier is all you need? In: International Conference on Learning Representations (2022). https://arxiv.org/abs/2110.06207

[10] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017)

[11] Tonmoy, M.T.H., Mahmud, S., Rahman, A.K.M.M., Amin, M.A., Ali, A.A.: Hierarchical self attention based autoencoder for open-set human activity recognition. In: Advances in Knowledge Discovery and Data Mining (PAKDD). Lecture Notes in Computer Science, vol. 12714, pp. 351–363 (2021). https://doi.org/10.1007/978-3-030-75768-7_28

[12] Cortese, A., Solbiati, S., Scandelli, A., Giudici, A., Antonello, N., Trojaniello, D., Boracchi, G., Caiani, E.G.: Open-set recognition of human activities from head-mounted inertial sensor. Sensors **26**(3), 1079 (2026) https://doi.org/10.3390/s26031079

[13] Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(3), 226–239 (1998) https://doi.org/10.1109/34.667881

[14] Wei, H., Jafari, R., Kehtarnavaz, N.: Fusion of video and inertial sensing for deep learning-based human action recognition. Sensors **19**(17), 3680 (2019) https://doi.org/10.3390/s19173680

[15] Ranieri, C.M., MacLeod, S., Dragone, M., Vargas, P.A., Romero, R.A.F.: Activity recognition for ambient assisted living with videos, inertial units and ambient sensors. Sensors **21**(3), 768 (2021) https://doi.org/10.3390/s21030768

[16] Dong, H., Zhao, Y., Chatzi, E., Fink, O.: MultiOOD: Scaling out-of-distribution detection for multiple modalities. In: Advances in Neural Information Processing Systems, vol. 37 (2024)

[17] Im, H., Lim, W., Kim, D.-W.: MAND: Modality-aware novelty detection for open-world egocentric activity recognition. In: EIML Workshop at ICML (2026). https://arxiv.org/abs/2603.16970

[18] Doorenbos, L., Vu, D.M., Ozsoy, S., Gall, J.: Modality-aware out-of-distribution detection for multi-modal action recognition. In: European Conference on Computer Vision (2026). Accepted paper. https://arxiv.org/abs/2606.24404

[19] Chen, C., Jafari, R., Kehtarnavaz, N.: UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: IEEE International Conference on Image Processing, pp. 168–172 (2015). https://doi.org/10.1109/ICIP.2015.7350781

[20] Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Proceedings of the 13th Scandinavian Conference on Image Analysis. Lecture Notes in Computer Science, vol. 2749, pp. 363–370 (2003). https://doi.org/10.1007/3-540-45103-X_50

[21] Cho, K., Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of EMNLP, pp. 1724–1734 (2014). https://doi.org/10.3115/v1/D14-1179

[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)