STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, December 2026, pp 5661–5684.
Published online in International Academic Press (www.IAPress.org)

Stability-Aware Video-Window-Wise Bayesian Optimization of CLAHE for Ordinal Drowsiness Classification in Near-Infrared Facial Video

Moh Hadi Subowo [1,2,*], Pulung Nurtantio Andono [1], Guruh Fajar Shidik [1], Heru Agus Santoso [1]

[1] *Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia*
[2] *Faculty of Science and Technology, Universitas Islam Negeri Walisongo, Semarang, Indonesia*

Abstract Contrast Limited Adaptive Histogram Equalization (CLAHE) can improve near-infrared (NIR) facial video, but the temporal granularity at which its parameters are optimized may affect both visual consistency and downstream sequence learning. This study reformulates CLAHE tuning as stability-aware video-window-wise Bayesian Optimization (BO), where one parameter pair is applied to all frames in a ten-second input window and the objective jointly represents perceptual quality and temporal consistency.

The analysis uses 2,140 DROZY facial-video windows from 14 participants and 36 usable sessions. Representation-level experiments show a systematic quality–stability trade-off: the quality-only objective attains the lowest normalized PIQE (0.3228), whereas SA 30:70 attains the highest tSSIM (0.7928) and lowest FDV (107.84).

The complete eight-condition classifier experiment was evaluated using strict leave-one-subject-out (LOSO) validation, producing 112 separate MobileNetV2–GRU LOSO training runs. Under pooled out-of-subject evaluation, SA 30:70 achieved the highest Accuracy (0.5949; 95% subject-cluster bootstrap CI 0.4976–0.6935) and QWK (0.3147; 0.1337–0.4897), Flatten/legacy Window-wise BO achieved the highest Macro-F1 (0.4580; 0.3675–0.5517), and Static CLAHE remained highly competitive (Accuracy 0.5921; Macro-F1 0.4375; QWK 0.3050).

Friedman tests indicated overall differences for subject-level Accuracy ($p = 0.0138$) and fixed-three-class Macro-F1 ($p = 0.0348$), but no pairwise comparison survived Holm correction. Paired two-one-sided equivalence tests showed that SA 30:70 is equivalent to Static CLAHE and to legacy Window-wise BO within ± 0.05 subject-level Accuracy and Macro-F1, while providing markedly higher temporal stability (tSSIM 0.793 versus 0.721; FDV 108 versus 311 relative to Static CLAHE). The pattern was retained with training-fold class weighting, which raised Neutral recall by 0.14–0.20 and made every leading CLAHE condition significantly better than Original after Holm correction, and with a MobileNetV2–LSTM classifier, where SA 30:70 obtained the highest pooled Accuracy (0.6112). Under an equal 13-evaluation budget, BO outperformed Random Search on 85–91% of windows and matched a structured grid. Stability-aware optimization required approximately 5.9 s per ten-second window on one CPU thread, versus 16.9 s for frame-wise BO.

The results support temporal stability as a preprocessing design dimension that can be raised substantially without reducing unseen-subject classification beyond the equivalence margin, while not establishing a universally superior CLAHE configuration.

Keywords Drowsiness Detection, CLAHE, Bayesian Optimization, Temporal Granularity, Stability-Aware Optimization, Leave-One-Subject-Out, Subject-Independent Evaluation, PIQE, tSSIM, FDV, MobileNetV2-GRU, Near-Infrared Video

AMS 2010 subject classifications 68T10, 68T45, 62H30

DOI: 10.19139/soic-2310-5070-4625

*Correspondence to: Moh Hadi Subowo (Email: p41202300053@mhs.dinus.ac.id). Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2026 International Academic Press

1. Introduction

Driver drowsiness is a safety-relevant state that can be monitored non-intrusively through facial video. Deep drowsiness systems increasingly combine spatial convolutional representations with temporal models, including CNN–LSTM, graph-based, and transformer-like architectures [16,18–20], as well as hybrid deep-feature, multi-CNN, and attention-based temporal models [17,23–25,27,28]. Their reliability can nevertheless degrade when facial imagery has low contrast or illumination is limited. The DROZY database is particularly useful for studying this problem because it provides NIR facial recordings together with session-level Karolinska Sleepiness Scale (KSS) annotations [1].

CLAHE is a common local contrast-enhancement technique whose behavior depends strongly on clip limit and tile-grid size. Earlier work has therefore investigated automated parameter tuning using population-based optimization, handcrafted adaptation, and image-quality objectives [2–5]. In driver monitoring, Auto-CLAHE and Bayesian-optimized CLAHE have been integrated with deep temporal models [6,7]. A preceding NIR enhancement study from this research compared histogram-equalization and Retinex families and motivated CLAHE as the enhancement method carried into the present optimization study [8].

The central question in this paper is not whether CLAHE or BO is new in isolation. Rather, the contribution is a preprocessing design principle: the optimization unit should be aligned with the temporal unit consumed by the sequence model. Independent frame-level tuning can improve a frame-wise perceptual objective while changing enhancement parameters inside a sequence. A video-window-wise formulation instead selects one parameter pair for the complete temporal input, and a stability-aware objective can explicitly trade perceptual quality against inter-frame consistency.

This positioning differs from prior Auto-CLAHE and Bayesian-optimized CLAHE studies, which primarily optimize image quality or classification-oriented preprocessing without explicitly making temporal stability a component of the per-window optimization objective [6,7]. Within *Statistics, Optimization & Information Computing* (SOIC), Alzami et al. [7] demonstrated Bayesian-optimized CLAHE integrated with a TimeDistributed MobileNetV2–GRU architecture for drowsiness detection, while Subowo et al. [9] mapped the broader development of facial-feature-based drowsiness detection and highlighted continuing generalization and deployment gaps. These SOIC studies provide direct context for the present work, which extends the discussion from adaptive enhancement and field-level trends toward explicit temporal-stability-aware, subject-independent preprocessing evaluation.

This study makes five contributions. First, it quantifies the difference between frame-wise and video-window-wise parameter granularity using harmonized PIQE and within-window parameter drift. Second, it formulates stability-aware video-window-wise BO that jointly searches CLAHE clip limit and tile-grid size while varying the relative weights of aggregate PIQE and temporal instability. Third, it validates all eight preprocessing conditions using strict subject-independent LOSO across 14 participants, making subject-independent evaluation the primary generalization evidence. Fourth, it integrates pooled out-of-subject metrics, subject-cluster bootstrap confidence intervals, participant-level and class-level error analysis, and corrected subject-level inference to distinguish descriptive rankings from statistically supported conclusions. Fifth, it tests the robustness of these conclusions with equivalence testing, training-fold class weighting, a second recurrent classifier, a budget-matched optimizer benchmark, normalization sensitivity, and method-by-method preprocessing runtime, so that practitioners can see what temporal stability costs and what it does not change.

Figure 1 summarizes the complete experimental path from the NIR source videos through temporal-granularity analysis, stability-aware optimization, strict LOSO validation, and the final generalization interpretation. This workflow is included to make explicit how the precursor experiment and the final subject-independent experiment form one continuous methodological sequence rather than separate studies.

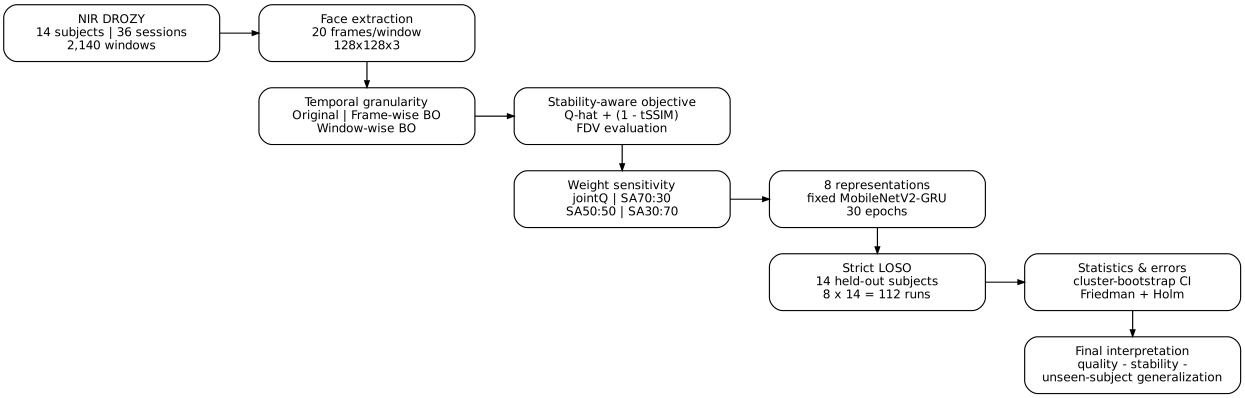


Figure 1. Overall experimental workflow from DROZY NIR videos to subject-independent LOSO interpretation. The pipeline separates the precursor temporal-granularity experiment, the final stability-aware objective study, and the downstream generalization analysis.

2. Related Work

2.1. CLAHE Parameter Optimization and Image-Quality Objectives

CLAHE adapts local contrast while limiting histogram amplification. Because clip limit and tile-grid size control the degree and spatial scale of enhancement, several studies have automated their selection. Adaptive and optimization-based methods include dynamic clip-limit selection, multi-objective Cuckoo Search, population-based optimization, and BO [2–5,7].

PIQE is attractive when no reference image is available [10]; however, optimizing a perceptual score on isolated images does not ensure that adjacent frames remain consistent. In the drowsiness domain, the SOIC study by Alzami et al. [7] is particularly relevant because it combines Bayesian-optimized CLAHE with a TimeDistributed MobileNetV2–GRU pipeline, providing a close methodological reference point for the present temporal-stability-aware reformulation.

BO is used as a sequential search mechanism for an expensive black-box per-window objective with a deliberately limited evaluation budget [12,13]. Random Search is a strong baseline when only a few dimensions matter [33], and structured designs can be competitive in very low-dimensional spaces. Because the present search space has only two parameters, the choice of BO is therefore tested empirically against Random Search and a structured grid under the same 13-evaluation budget (Section 5.7) rather than assumed.

2.2. Temporal Modeling in Drowsiness Detection

Sequence-based drowsiness detection uses temporal dependencies in facial behavior rather than treating each frame as independent. CNN–RNN, graph convolutional, and attention-based models have been used for this purpose [16,18–20].

Video-processing literature also shows that frame-independent transformations can introduce temporally inconsistent appearance [21,22]. This motivates evaluating preprocessing not only through a spatial quality metric but also through temporal consistency.

Temporal consistency has also been imposed during learning, for example by regularizing single-frame networks to respond stably to small temporal perturbations [40]. Self-supervised video representation learning exploits the same property from the opposite direction: slowly varying appearance, temporal order, and spatiotemporal contrastive objectives provide supervisory signal precisely because neighbouring frames should map to nearby representations [37–39]. Sequence classifiers for drowsiness rely on this regularity when they model eyelid, mouth, and head dynamics across a window [16,18–20,25]. The present study applies the idea one step earlier in the pipeline: instead of asking the network to tolerate unstable inputs, it asks the preprocessing optimizer to avoid creating them.

A parallel line of work evaluates drowsiness systems under deployment constraints, including low-cost embedded hardware [29], TinyML models [30], IoT alerting pipelines [26,27], and recent real-time or multimodal systems [31,32]. These studies make computational cost part of the evaluation, and the bibliometric analysis in [9] found that operational metrics such as FPS, latency, and time-to-alarm are rarely reported. A recent SOIC review of driver monitoring systems likewise identifies heavy reliance on laboratory or simulator data and the need for computationally efficient algorithms [41]. The present revision therefore reports method-by-method preprocessing runtime (Section 5.9) even though full end-to-end deployment is outside its scope.

In this study, the ordered Alert–Neutral–Sleepy labels are modeled during training with a three-class softmax classifier so that the downstream model remains identical across preprocessing conditions. Ordinal structure is incorporated at evaluation through Quadratic Weighted Kappa (QWK) and through explicit analysis of class-confusion behavior. We do not claim that nominal softmax is superior to ordinal regression; ordinal-aware losses constitute a separate modeling factor that would confound the present effort to isolate preprocessing effects. Training-fold class weighting, which was part of the original study design, is evaluated as a sensitivity analysis (Section 5.6).

2.3. Positioning of the Present Study

The novelty is therefore methodological rather than architectural: CLAHE optimization is reformulated from independent spatial decisions into a video-window-wise decision whose objective explicitly contains a temporal-stability term.

The strongest claim is correspondingly limited to this preprocessing design principle and to the evaluated DROZY setting with MobileNetV2-based recurrent classifiers (GRU as the primary model and LSTM as a sensitivity model). Cross-dataset generalization and non-recurrent temporal architectures are not inferred from the present experiment.

3. Method

3.1. Dataset, Labels, Window Construction, and Subject Lineage

The study uses 36 usable NIR sessions from 14 DROZY participants [1]. Ten-second windows contain 20 sampled face frames (2 frames/s), each resized to $128 \times 128 \times 3$.

KSS values are grouped into three ordered states: Alert ($\mathrm{KSS} \leq 3$), Neutral ($4 \leq \mathrm{KSS} \leq 5$), and Sleepy ($\mathrm{KSS} \geq 6$).

The final window counts are 658 Alert, 293 Neutral, and 1,189 Sleepy, for 2,140 windows in total.

The feature arrays used in the initial experiments did not store subject identifiers directly. For the subject-independent analysis, subject/session lineage was reconstructed from the original source-video traversal used by the data-generation pipeline and subsequently verified against the generated arrays.

Every one of the 2,140 windows received one verified subject and session assignment. This mapping is used only to define subject-independent splits; no label, image representation, or prediction was altered during the lineage reconstruction.

The resulting subject structure is heterogeneous. Two participants contain only one observed ordinal class, while several others contain two of the three classes. This property is important for the interpretation of subject-level metrics: Accuracy and fixed-three-class Macro-F1 can be compared across all 14 held-out participants, whereas participant-level QWK is not a suitable paired inferential statistic when the true labels of a held-out participant contain a single class.

Figure 2 visualizes the class composition of each held-out participant, and Table 1 reports the exact subject/session/window structure used by the LOSO protocol.

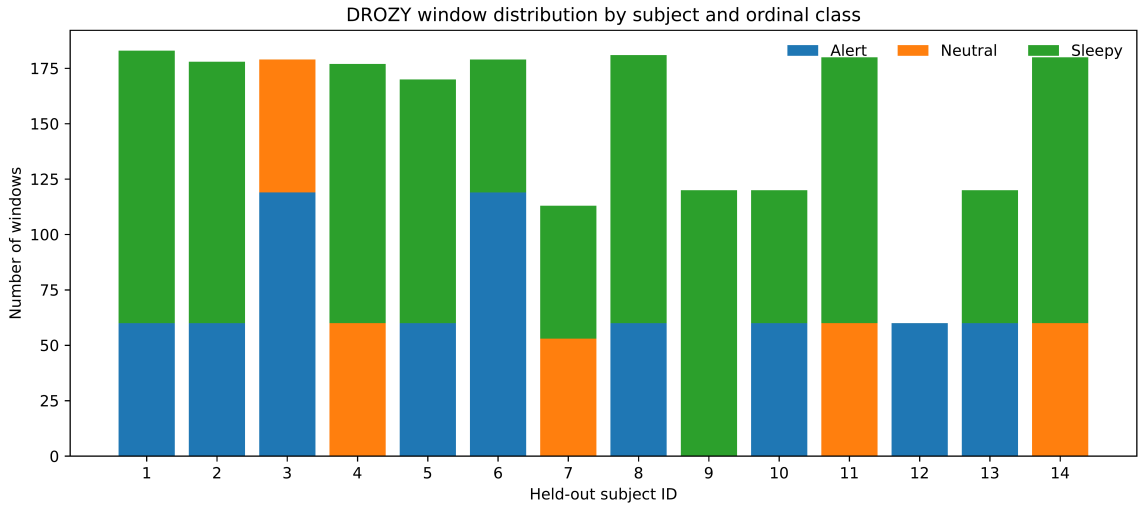


Figure 2. Distribution of the 2,140 windows across the 14 DROZY participants and three ordinal classes. The heterogeneous class composition motivates fixed-three-class Macro-F1 and explains why subject-level QWK is not used for paired inference.

Table 1. Verified subject/session structure used for strict LOSO.

Subject	Sessions	Alert	Neutral	Sleepy	Windows	Classes present
1	3	60	0	123	183	Alert, Sleepy
2	3	60	0	118	178	Alert, Sleepy
3	3	119	60	0	179	Alert, Neutral
4	3	0	60	117	177	Neutral, Sleepy
5	3	60	0	110	170	Alert, Sleepy
6	3	119	0	60	179	Alert, Sleepy
7	2	0	53	60	113	Neutral, Sleepy
8	3	60	0	121	181	Alert, Sleepy
9	2	0	0	120	120	Sleepy only
10	2	60	0	60	120	Alert, Sleepy
11	3	0	60	120	180	Neutral, Sleepy
12	1	60	0	0	60	Alert only
13	2	60	0	60	120	Alert, Sleepy
14	3	0	60	120	180	Neutral, Sleepy
Total	36	658	293	1189	2140	–

3.2. Common Face Extraction and Window Representation

The preprocessing comparison begins from a common face representation. For each source video, a frontal-face Haar Cascade detector identifies the face region; the selected full-face bounding box is cropped, resized to 128×128 pixels, and represented in three channels. Twenty detected face frames are retained for each ten-second input window. Consequently, all eight experimental conditions share the same window count, temporal length, spatial resolution, labels, and downstream architecture; only the image-enhancement strategy differs.

This common representation is important because it prevents changes in face localization, crop geometry, or sequence length from becoming hidden confounders in the comparison between Original, Static, Frame-wise, legacy Window-wise, and the four final joint-optimization conditions.

3.3. Temporal-Granularity Baselines and Harmonized PIQE Evaluation

Three precursor conditions are compared under a common post-hoc PIQE implementation: Original, Frame-wise BO, and legacy Window-wise BO.

Frame-wise BO optimizes clip limit independently for each of the 20 frames, with tile size fixed at 8×8. When the enhanced candidate does not improve the legacy PIQE criterion, the Original frame is retained.

Legacy Window-wise BO selects one clip limit for the entire window using three representative frames and applies that value to all 20 frames. The harmonized PIQE re-evaluation is descriptive and does not retroactively change the parameters originally selected by these legacy pipelines.

Within-window Frame-wise parameter drift is quantified from the 20 selected clip-limit values using their standard deviation and range. Window-wise within-window drift is zero by construction because one clip limit is applied to all 20 frames.

Figure 3 shows actual frames from a representative experiment window.

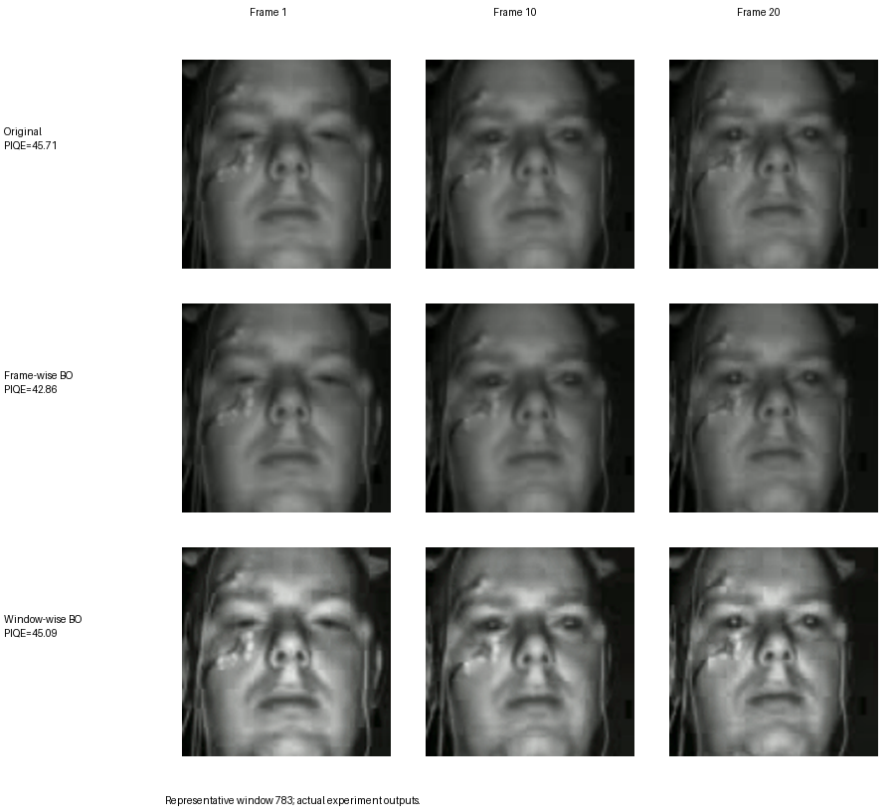


Figure 3. Actual experiment frames illustrating Original, Frame-wise BO, and legacy Window-wise BO on the same representative window.

3.4. Stability-Aware Joint Bayesian Optimization

For the final formulation, one CLAHE parameter pair $\theta = (c, r)$ is used for the entire window, where c is clip limit and r is tile-grid size:

$$c \in [0.1, 8.0], \qquad r \in \{4, 5, 6, 7, 8\}. \tag{1}$$

For window $W_k = \{f_{k,1}, \ldots, f_{k,T}\}$ with $T = 20$, aggregate quality is the mean PIQE over all enhanced frames:

$$Q(W_k, \theta) = \frac{1}{T} \sum_{t=1}^{T} \mathrm{PIQE}\left(E(f_{k,t}; \theta)\right), \qquad \widehat{Q} = \frac{Q}{100}. \tag{2}$$

Temporal structural similarity uses SSIM [11] over the 19 consecutive-frame pairs:

$$\mathrm{tSSIM}(W_k,\theta)=\frac{1}{T-1}\sum_{t=1}^{T-1}\mathrm{SSIM}(g_t,g_{t+1}),\qquad \widehat{S}=1-\mathrm{tSSIM}. \tag{3}$$

The scalarized objective minimized by BO is

$$J(W_k,\theta)=w_Q\widehat{Q}(W_k,\theta)+w_T\widehat{S}(W_k,\theta),\qquad w_Q+w_T=1. \tag{4}$$

Frame Difference Variance (FDV) remains evaluation-only:

$$\mathrm{FDV}(W_k)=\frac{1}{T-1}\sum_{t=2}^{T}\mathrm{Var}\left(|g_t-g_{t-1}|\right). \tag{5}$$

Four objectives are evaluated:

- jointQ $(1.0, 0.0)$,
- SA 70:30 $(0.7, 0.3)$,
- SA 50:50 $(0.5, 0.5)$,
- SA 30:70 $(0.3, 0.7)$.

A Gaussian-process surrogate with Matérn $\nu = 2.5$ and Expected Improvement with $\xi = 0.01$ is used [12,13]. Each window receives three initial evaluations followed by ten BO iterations, yielding 13 evaluated candidates.

The tile-grid dimension is evaluated only at the valid integer values $r \in \{4, 5, 6, 7, 8\}$. When the optimizer proposes the continuous representation used internally by the surrogate, the tile coordinate is mapped to the nearest valid integer before CLAHE evaluation, and the objective supplied back to the Gaussian process is always the score of that actually evaluated discrete configuration.

Thus, the Gaussian process is not interpreted as a categorical model; it is a surrogate over the mixed search representation whose objective calls are restricted to valid CLAHE configurations.

The normalization $\widehat{Q} = Q/100$ is an operational scaling rather than a theoretically universal calibration. Across the final selected V3 windows, observed raw PIQE values span approximately 11.15–54.73, depending on objective setting. Over all candidates evaluated in the optimizer benchmark (Section 5.7), the central 90% of raw PIQE lies in 28.9–47.9 and that of $1 - \mathrm{tSSIM}$ in 0.159–0.442, so the two scaled terms vary over ranges of similar size (0.19 versus 0.28). Dividing by 100 therefore places the quality component on a magnitude broadly comparable with the bounded instability term $1 - \mathrm{tSSIM}$. Section 5.8 tests whether min–max normalization would change the selected parameters.

Application to another dataset, camera, or perceptual metric may require recalibration or a calibration-free multi-objective formulation. Accordingly, no dataset-independent transferability claim is made for the numeric scaling itself.

Table 2 consolidates the exact BO configuration used by the final V3 conditions. The search budget is intentionally fixed at 13 evaluated candidates per window so that all weighted objectives have the same optimization budget and differ only in the scalarized objective.

3.5. Classification Architecture and Ordinal Evaluation

All preprocessing conditions use the same TimeDistributed MobileNetV2–GRU classifier.

MobileNetV2 uses width multiplier $\alpha = 0.5$, ImageNet initialization, and `include_top=False`; the last 30 backbone layers are trainable.

TimeDistributed global average pooling converts each frame to a feature vector. A GRU with 32 units is followed by dropout 0.2, Dense(16, ReLU), dropout 0.2, and Dense(3, softmax) [14,15].

All models use SGD with learning rate 0.001, sparse categorical cross-entropy, batch size 64, and 30 fixed epochs.

No early stopping, class weighting, or resampling is used in the primary experiment so that preprocessing remains the manipulated experimental factor. Two sensitivity variants keep every other setting fixed. The first

Table 2. Final V3 Bayesian-optimization configuration.

Component	Setting
Optimization unit	one 20-frame video window
Clip-limit domain	$[0.1, 8.0]$
Tile-grid domain	integer $\{4, 5, 6, 7, 8\}$
Gaussian-process kernel	Matérn $\nu = 2.5$
Acquisition	Expected Improvement
Exploration parameter	$\xi = 0.01$
Initial evaluations	3
BO iterations	10
Total evaluations/window	13
Quality term	$\widehat{Q} = \mathrm{mean}(\mathrm{PIQE})/100$
Stability term	$\widehat{S} = 1 - \mathrm{tSSIM}$
Independent evaluation	FDV
Randomization	deterministic window-specific seed

Table 3. Preprocessing conditions evaluated by the downstream classifier.

Condition	Granularity	Search/setting	Objective / role
Original	Frame	Linear contrast stretch	Reference representation
Static CLAHE	Fixed	clip=2.0, tile=8×8	Fixed CLAHE baseline
Frame-wise BO	Frame	clip $[0.1, 2.0]$, tile=8	PIQE per frame + fallback
Flatten/Window-wise BO	Window	clip $[0.1, 2.0]$, tile=8	Three-frame mean PIQE
jointQ	Window	clip $[0.1, 8.0]$, tile 4–8	$J = \widehat{Q}$
SA 70:30	Window	final domain	$0.7\widehat{Q} + 0.3\widehat{S}$
SA 50:50	Window	final domain	$0.5\widehat{Q} + 0.5\widehat{S}$
SA 30:70	Window	final domain	$0.3\widehat{Q} + 0.7\widehat{S}$

adds balanced class weights $w_c = N/(3N_c)$ computed only from the 13 training participants of each fold. The second replaces GRU(32) with LSTM(32). Weighted losses and the GRU–LSTM choice have been examined for imbalanced facial sequence data in a related SOIC study [42].

3.6. Software, Hardware, and Reproducibility Artifacts

The final LOSO runs use Python 3.11.10, TensorFlow 2.18.0, Keras 3.6.0, and an NVIDIA A100 80 GB GPU environment with cuDNN 9.3. For each run, the training history, test predictions, per-class metrics, confusion matrix, seed, split membership, runtime, and manifest metadata were retained. The final analysis package additionally retains pooled predictions, subject-level metrics, cluster-bootstrap summaries, and the executed analysis notebook. These artifacts allow the reported tables and figures to be regenerated without reconstructing the model-training stage.

4. Experimental Setup

4.1. Primary Strict Leave-One-Subject-Out Protocol

The primary classification evaluation uses strict leave-one-subject-out (LOSO) validation.

For each of 14 folds, all windows from one participant are held out completely for testing, while the remaining 13 participants are used for training.

No validation subject is introduced, no early stopping or model selection is performed, and the same fixed 30-epoch schedule is used in every fold. Fold-specific deterministic seeds are used consistently across preprocessing conditions.

The eight preprocessing conditions therefore produce

$$8 \times 14 = 112$$

independent training runs.

The pooled out-of-subject result for each condition is obtained by concatenating the held-out predictions from all 14 LOSO folds. This guarantees that every one of the 2,140 windows appears exactly once in a test set and that the model predicting a given window was never trained on that participant.

The split is paired across preprocessing conditions: for a given test subject, every condition uses the same held-out subject, the same training subjects, the same class labels, and the same fold-specific seed. The training-set size therefore ranges only according to the number of windows belonging to the held-out subject, from 1,957 to 2,080 training windows. This paired construction isolates the representation choice while keeping the subject-generalization task fixed.

Subject-level descriptive metrics are retained for all 14 held-out participants.

Session-level metrics are additionally summarized descriptively from the same out-of-subject predictions. They are not treated as independent inferential units because sessions are nested within participants and individual sessions contain one dominant label.

4.2. Secondary Window-Level Evaluation

For comparison, the earlier evaluation protocol used a 70/30 window-level stratified partition. Five fold-specific models were trained within the 70% training pool and evaluated on the same 642-window held-out test set.

Those results are retained only as a secondary comparison for quantifying the generalization gap. They are not interpreted as evidence of unseen-subject performance because the protocol did not explicitly enforce participant or session separation between training and testing.

4.3. Statistical Analysis

For each preprocessing condition, pooled Accuracy, three-class Macro-F1, and QWK are computed from the complete out-of-subject predictions.

Ninety-five percent confidence intervals are estimated using 2,000 subject-cluster bootstrap iterations, resampling participants rather than individual windows. This preserves within-participant dependence.

For inferential comparison, the 14 held-out participants form paired blocks. Friedman tests are applied to subject-level Accuracy and fixed-three-class Macro-F1.

When an omnibus difference is detected, paired two-sided Wilcoxon signed-rank tests are used for pairwise contrasts and Holm correction is applied across all comparisons. Rank-biserial correlation is retained as an effect-size descriptor.

Subject-level QWK is not used in Friedman/Wilcoxon inference because some held-out participants contain only one true class, making kappa undefined or non-informative at that level.

QWK is therefore reported as a pooled metric with subject-cluster bootstrap confidence intervals.

4.4. Robustness Protocols

Equivalence. Because absence of a significant difference is not evidence of equivalence, paired two-one-sided tests (TOST) [35] are applied to subject-level Accuracy and Macro-F1 ($n = 14$). Two conditions are declared equivalent when the 90% confidence interval of the paired difference lies inside $\pm\delta$; $\delta = 0.05$ is the primary margin and $\delta = 0.10$ a secondary one.

Class weighting and recurrent unit. The class-weighted GRU is evaluated with strict LOSO for all eight conditions (112 runs). The LSTM variant is evaluated with strict LOSO for five representative conditions (Original, Static CLAHE, legacy Window-wise BO, jointQ, and SA 30:70; 70 runs). Both reuse the canonical subject splits and fold-specific seeds, so each run is paired with its canonical counterpart. Holm correction [34] is applied within each family of comparisons.

Optimizer benchmark. On 140 windows stratified by participant and class, BO with the exact canonical settings is compared with Random Search (uniform clip limit and tile size, three repeats averaged) and a deterministic space-filling structured grid of 13 points over $c \in [0.1, 8.0]$ and $r \in \{4, \ldots, 8\}$. Every method receives exactly 13 objective calls per window; repeated BO proposals count against the budget, as in the canonical generator. Regret is the difference between the best-so-far objective and the best value found by any method for that window.

Normalization sensitivity. Using the union of candidates evaluated in the benchmark, the parameter pair selected under $\widehat{Q} = Q/100$ is compared with the pair selected under per-window min–max scaling of both terms, dataset-range min–max scaling, and robust 5th–95th percentile scaling, for all four weightings.

Window-level association. Harmonized PIQE, tSSIM, and FDV are recomputed for every window of all eight representations and joined with the strict-LOSO predictions. Spearman correlations describe marginal association with correctness and with the predicted probability of the true class. A subject-clustered generalized estimating equation (GEE) [36] with logit link, exchangeable working correlation, standardized metrics, and condition fixed effects tests within-condition association.

Runtime. Preprocessing time per 20-frame window (parameter search plus application) is measured for all eight conditions on 40 stratified windows with two repeats, interleaving conditions, on a single CPU thread. Face detection and cropping, common to all conditions, are excluded. Frame-wise and legacy Window-wise BO are re-implemented from the protocol in Section 3.3 with three initial points and ten iterations, solely for timing.

5. Results

5.1. Temporal Granularity as a Precursor Experiment

The harmonized PIQE comparison shows that Frame-wise BO succeeds at its local quality objective.

Mean PIQE falls from 45.7657 in Original to 40.3444 in Frame-wise BO and 42.1547 in legacy Window-wise BO.

Frame-wise BO is lower than Original on 2,107 of 2,140 windows (98.46%) and lower than Window-wise BO on 1,794 windows (83.83%).

Friedman testing gives

$$\chi^2 = 2819.431, \qquad p < 0.0001,$$

with Kendall's $W = 0.659$; all three paired contrasts remain significant after Holm adjustment.

The parameter-level behavior differs substantially. Frame-wise BO changes clip limit inside every analyzed window, with mean within-window standard deviation 0.5466.

CLAHE is selected on 27,307 of 42,800 frames (63.80%), while 15,493 frames (36.20%) fall back to Original.

Legacy Window-wise BO applies one clip limit to the complete 20-frame window and therefore has zero within-window parameter drift by construction.

These observations support a mechanism-level association rather than a causal mediation claim. The present study does not directly measure how CLAHE parameter changes propagate into MobileNetV2 embeddings or GRU state trajectories.

5.2. Quality–Stability Sensitivity of the Final Objective

The final objective behaves systematically as temporal weight increases.

jointQ produces the lowest $\widehat{Q}$ but also the lowest tSSIM and highest FDV.

Moving through SA 70:30 and SA 50:50 toward SA 30:70 increases structural consistency and decreases frame-difference variance while accepting a less favorable PIQE-oriented quality score.

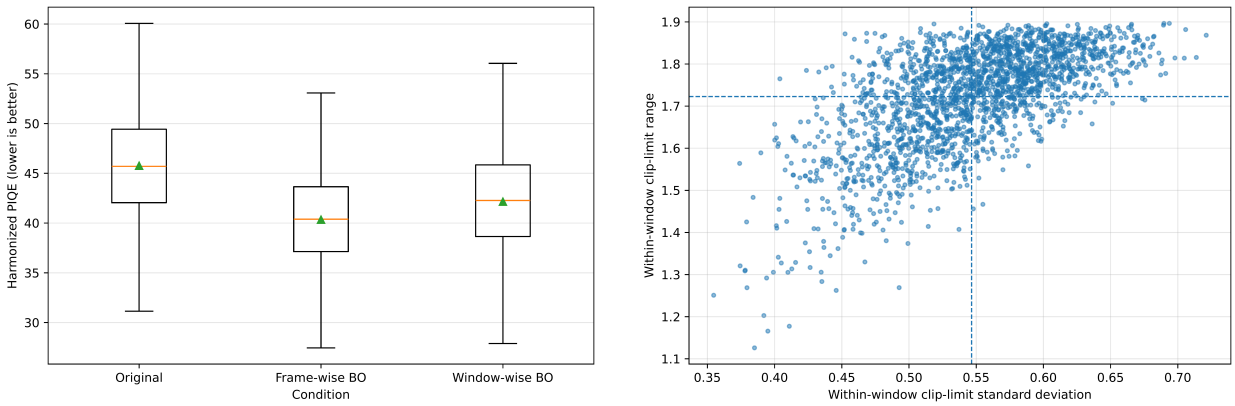


Figure 4. Temporal-granularity evidence. Left: harmonized PIQE distributions across 2,140 paired windows. Right: within-window Frame-wise clip-limit drift.

Table 4. Final preprocessing metrics across all 2,140 windows (mean ± standard deviation).

Condition	$\widehat{Q}\downarrow$	tSSIM ↑	FDV ↓
Static CLAHE	0.3985±0.0506	0.7209±0.0829	311.20±161.08
jointQ	**0.3228±0.0495**	0.6837±0.0874	504.81±243.19
SA 70:30	0.3295±0.0530	0.6997±0.0886	444.27±246.55
SA 50:50	0.3667±0.0659	0.7521±0.0782	243.39±200.05
SA 30:70	0.4166±0.0524	**0.7928±0.0713**	**107.84±83.73**

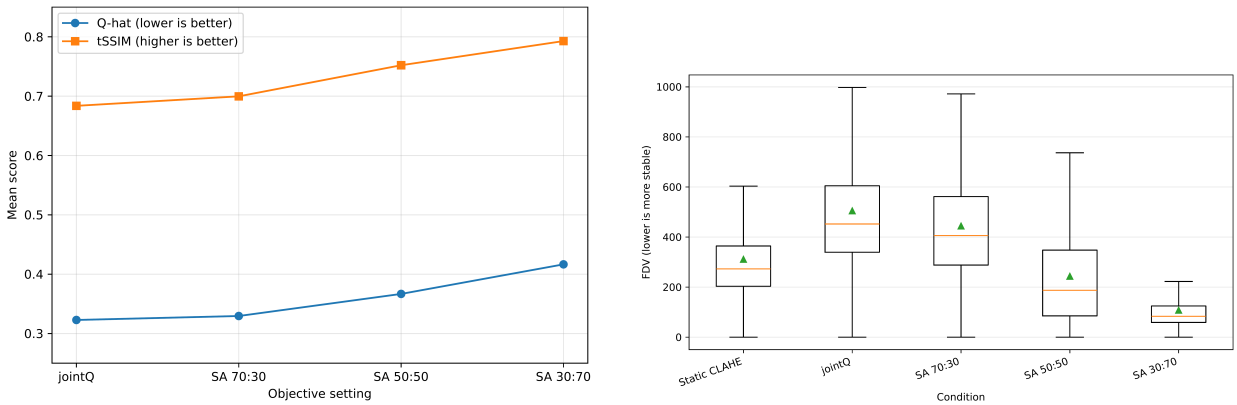


Figure 5. Quality–stability sensitivity. Left: mean $\widehat{Q}$ and tSSIM across objective settings. Right: FDV distributions for Static CLAHE and the four adaptive final objectives.

The representation-level differences are statistically large at the paired window level. Table 5 reports Friedman omnibus tests across Static CLAHE and the four V3 conditions. These tests characterize systematic within-corpus differences among preprocessing outputs; they are not used as evidence of subject-independent classification superiority.

Table 5. Friedman tests across Static CLAHE and the four V3 preprocessing conditions ($n = 2,140$ paired windows).

Metric	χ^2	p	Kendall's W
$\widehat{Q}$	5739.341	< 0.0001	0.670
tSSIM	6003.065	< 0.0001	0.701
FDV	5888.578	< 0.0001	0.688

The two temporal measures are also internally consistent. Across all five V3 conditions, tSSIM and FDV are strongly and inversely associated, with Spearman coefficients from -0.805 to -0.899. By contrast, the relationship

between $\widehat{Q}$ and either stability measure is weaker and condition dependent (Table 6). Thus, the quality and stability terms capture related but non-identical properties of the enhanced windows.

Table 6. Spearman relationships among preprocessing metrics across all 2,140 windows.

Condition	$\rho(\widehat{Q}, \mathrm{tSSIM})$	$\rho(\widehat{Q}, \mathrm{FDV})$	$\rho(\mathrm{tSSIM}, \mathrm{FDV})$
Static CLAHE	0.260	-0.369	-0.898
jointQ	0.114	-0.201	-0.899
SA 70:30	0.222	-0.349	-0.882
SA 50:50	0.374	-0.651	-0.805
SA 30:70	0.226	-0.259	-0.866

5.3. Strict LOSO Out-of-Subject Classification

Table 7 reports the primary subject-independent results.

SA 30:70 obtains the highest pooled Accuracy (0.5949) and QWK (0.3147), while Flatten/legacy Window-wise BO obtains the highest pooled Macro-F1 (0.4580).

Static CLAHE is nearly tied in pooled Accuracy (0.5921) and remains competitive on Macro-F1 and QWK.

Original is substantially lower on all three pooled metrics.

Table 7. Strict LOSO pooled out-of-subject classification with 95% subject-cluster bootstrap confidence intervals. Bold values indicate the highest point estimate in each column, not statistical superiority.

Condition	Accuracy [95% CI]	Macro-F1 [95% CI]	QWK [95% CI]
Original	0.4290 [0.3288, 0.5193]	0.2896 [0.2250, 0.3437]	-0.0678 [-0.2557, 0.1104]
Static CLAHE	0.5921 [0.4918, 0.6957]	0.4375 [0.3731, 0.5081]	0.3050 [0.1309, 0.4784]
Frame-wise BO	0.5477 [0.4420, 0.6544]	0.3789 [0.3034, 0.4535]	0.2059 [0.0010, 0.4017]
Flatten/Window-wise BO	0.5916 [0.4863, 0.7085]	**0.4580 [0.3675, 0.5517]**	0.3016 [0.1170, 0.4927]
jointQ	0.5523 [0.4451, 0.6434]	0.3985 [0.3235, 0.4596]	0.2130 [0.0601, 0.3584]
SA 70:30	0.5364 [0.4384, 0.6260]	0.3693 [0.3080, 0.4211]	0.1852 [0.0484, 0.3270]
SA 50:50	0.5682 [0.4671, 0.6674]	0.3915 [0.3351, 0.4501]	0.2542 [0.1049, 0.4198]
SA 30:70	**0.5949 [0.4976, 0.6935]**	0.4484 [0.3616, 0.5356]	**0.3147 [0.1337, 0.4897]**

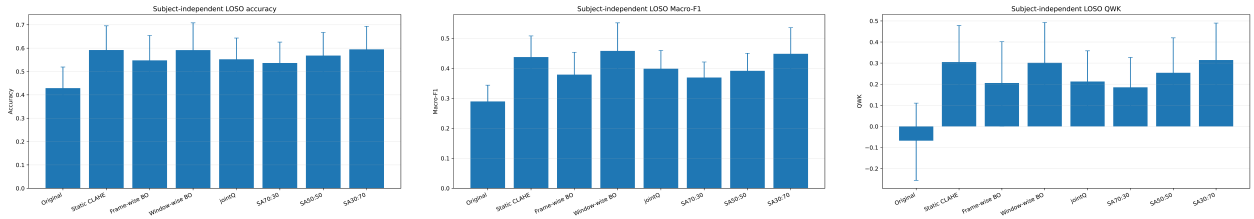


Figure 6. Pooled strict-LOSO Accuracy, Macro-F1, and QWK across eight preprocessing conditions.

Subject-level Friedman tests indicate overall differences for Accuracy

$$\chi^2 = 17.623, \qquad p = 0.0138, \qquad W = 0.180,$$

and fixed-three-class Macro-F1

$$\chi^2 = 15.098, \qquad p = 0.0348, \qquad W = 0.154.$$

However, none of the 56 pairwise Wilcoxon comparisons across these two metrics remains significant after Holm correction.

Table 8 summarizes the inferential boundary. The unadjusted contrasts against Original often have large directional rank-biserial effects, but the multiplicity-adjusted evidence is insufficient to label any individual pair as statistically superior. For example, the Accuracy contrast between Original and Static has rank-biserial magnitude 0.872 but Holm-adjusted $p = 0.199$; Original versus SA 30:70 has magnitude 0.846 with adjusted $p = 0.193$.

Table 8. Subject-level inferential summary under strict LOSO. Negative rank-biserial values indicate lower performance of the first condition.

Analysis	Contrast	χ^2/stat.	Raw p	Holm p	Rank-biserial
Friedman Accuracy	8 conditions	17.623	0.0138	–	$W = 0.180$
Friedman Macro-F1	8 conditions	15.098	0.0348	–	$W = 0.154$
Wilcoxon Accuracy	Original–Static	5.0	0.00765	0.1989	-0.872
Wilcoxon Accuracy	Original–Flatten	9.0	0.01075	0.2687	-0.802
Wilcoxon Accuracy	Original–SA 30:70	7.0	0.00713	0.1926	-0.846
Wilcoxon Accuracy	Static–SA 30:70	38.0	0.93747	1.0000	-0.026

The descriptive ranking in Table 7 must therefore not be interpreted as pairwise statistical superiority.

Among the three weightings that include a stability term, the pooled task metrics increase with the temporal-stability weight.

Accuracy rises from 0.5364 (SA 70:30) to 0.5682 (SA 50:50) and 0.5949 (SA 30:70).

Macro-F1 rises from 0.3693 to 0.3915 and 0.4484, while QWK rises from 0.1852 to 0.2542 and 0.3147.

This ordering does not extend to the quality-only end of the range: jointQ ($w_T = 0$) is slightly higher than SA 70:30 on all three pooled metrics (Accuracy 0.5523, Macro-F1 0.3985, QWK 0.2130). Across the full weight range the response therefore dips at $w_T = 0.3$ before rising, and with three or four points and overlapping intervals it is reported only as a descriptive pattern. The same shape recurs under class weighting (Section 5.6).

5.4. Participant-, Session-, and Class-Level Behavior

Figure 7 shows substantial heterogeneity across held-out participants together with a consistent class-specific difficulty.

Because subject class composition differs, participant Accuracy should not be interpreted as a pure ranking of participant difficulty. Nevertheless, the heatmap demonstrates that performance variation is not driven by one uniformly easy or difficult participant.

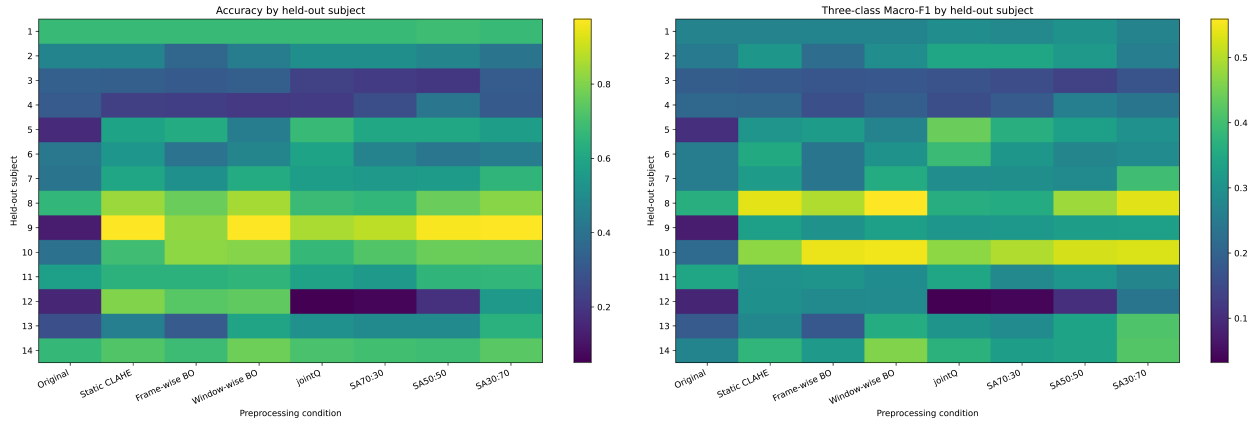


Figure 7. Strict-LOSO participant-level heterogeneity. Left: Accuracy by held-out participant and preprocessing condition. Right: fixed-three-class Macro-F1 by held-out participant and condition.

Neutral is the principal bottleneck.

Even the best Neutral recall, obtained by Flatten/Window-wise BO, is only 0.1263, compared with 0.4833 for Alert and 0.7662 for Sleepy in the same condition.

SA 30:70 recalls 0.4483 of Alert, 0.1092 of Neutral, and 0.7956 of Sleepy.

The class imbalance (293 Neutral windows versus 658 Alert and 1,189 Sleepy) is one contributor, while the intermediate ordinal position of Neutral may also create visual overlap with both neighboring states. The latter interpretation is plausible but is not directly isolated by the present experiment.

Table 9 makes the class imbalance visible at the outcome level. Flatten has the highest Alert and Neutral recall among the three top operating points, whereas SA 30:70 has the highest Sleepy recall. The main limitation across

all methods is therefore not Sleepy recognition, but recovering the intermediate Neutral state without collapsing it into either neighboring ordinal category.

Table 9. Pooled strict-LOSO class recall for the principal operating points and the unenhanced baseline.

Condition	Alert	Neutral	Sleepy	Macro-F1
Original	0.1854	0.0546	0.6560	0.2896
Static CLAHE	0.4468	0.0853	0.7973	0.4375
Flatten/Window-wise BO	**0.4833**	**0.1263**	0.7662	**0.4580**
SA 30:70	0.4483	0.1092	**0.7956**	0.4484

Figure 8 provides the corresponding pooled confusion matrices. The reduction of large Alert-to-Sleepy errors relative to Original is visible for the three leading preprocessing strategies, while Neutral remains the smallest diagonal component in every matrix.

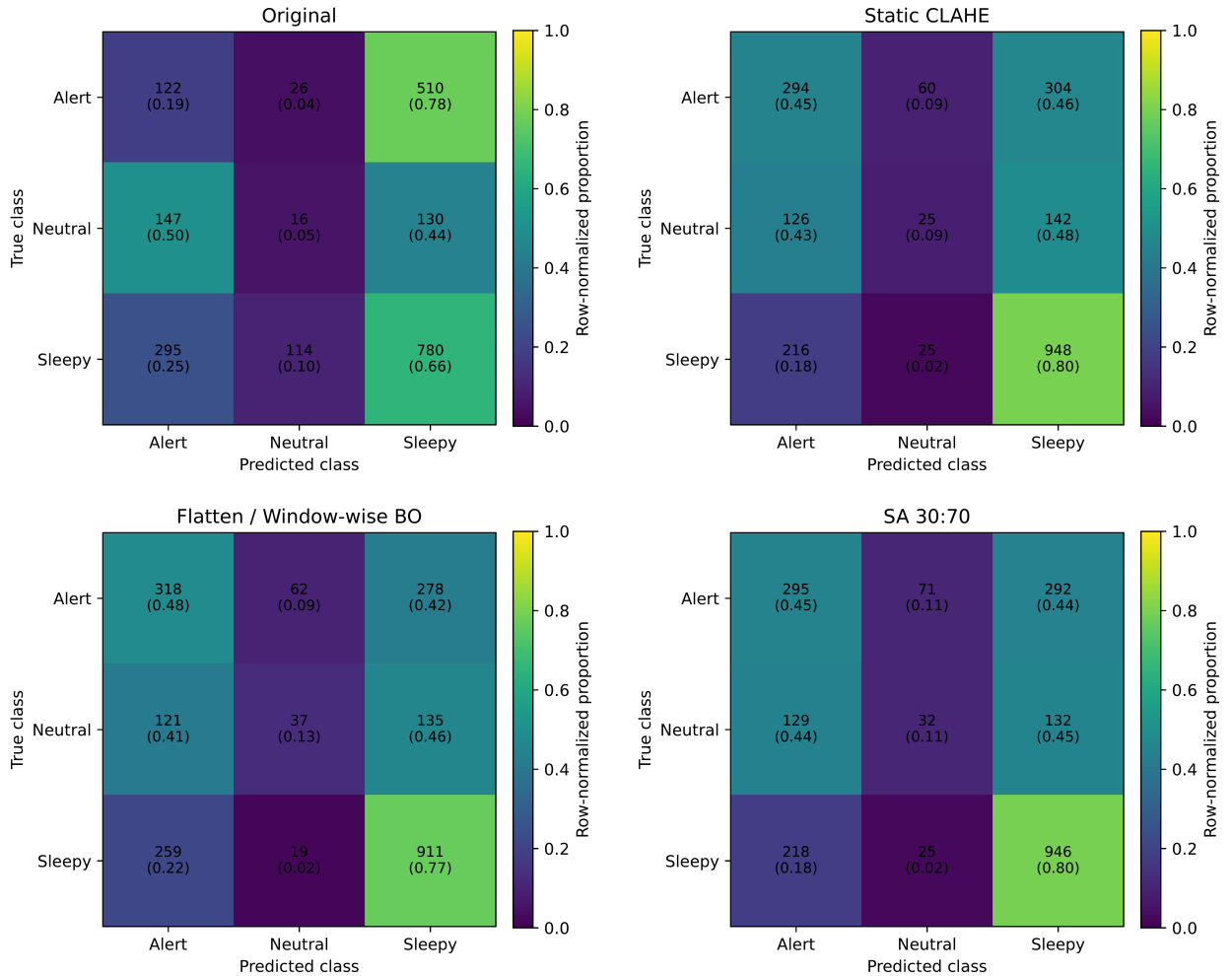


Figure 8. Pooled strict-LOSO confusion matrices for Original, Static CLAHE, Flatten/legacy Window-wise BO, and SA 30:70. Each cell shows the count and row-normalized proportion.

Descriptive session-level aggregation across the 36 sessions yields the same broad ordering: mean session Accuracy is 0.5952 for SA 30:70, 0.5920 for Static, 0.5909 for Flatten, and 0.4271 for Original.

Session-level QWK is not informative because individual sessions contain one dominant class; these session summaries are therefore descriptive only.

5.5. *Generalization Gap Relative to the Window-Level Evaluation*

The strict LOSO results are markedly lower than the earlier window-level estimates, demonstrating that the two protocols answer different questions.

Table 10 retains the earlier results for transparency but does not treat them as unseen-subject evidence.

Table 10. Window-level evaluation versus strict LOSO. Values are task metrics under different evaluation protocols and should not be treated as paired estimates of the same estimand.

Condition	Window Acc.	LOSO Acc.	Window Macro-F1	LOSO Macro-F1	Window QWK	LOSO QWK
Original	0.9421	0.4290	0.9177	0.2896	0.9168	-0.0678
Static CLAHE	0.9442	0.5921	0.9040	0.4375	0.9391	0.3050
Frame-wise BO	0.9159	0.5477	0.8685	0.3789	0.8891	0.2059
Flatten/Window-wise BO	0.9523	0.5916	0.9186	0.4580	0.9471	0.3016
jointQ	0.8988	0.5523	0.8169	0.3985	0.8781	0.2130
SA 70:30	0.8919	0.5364	0.8144	0.3693	0.8632	0.1852
SA 50:50	0.8798	0.5682	0.8058	0.3915	0.8330	0.2542
SA 30:70	0.9436	0.5949	0.9051	0.4484	0.9320	0.3147

Figure 9 visualizes the same protocol shift for all three primary metrics. The dominant pattern is a downward displacement under strict LOSO, but the size of the gap differs by representation. This figure is not interpreted as a paired statistical test because the two protocols estimate different targets; it is included to make the change in generalization question visually explicit.

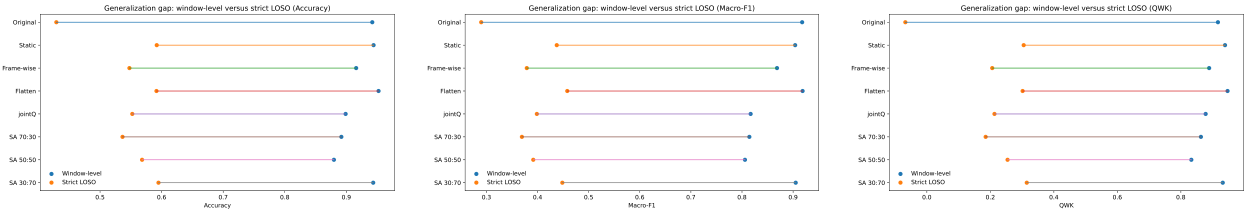


Figure 9. Descriptive generalization gap between the earlier window-level protocol and strict LOSO for Accuracy, Macro-F1, and QWK.

The largest qualitative change concerns the Original representation.

It appears competitive under the window-level protocol but becomes the lowest-performing condition under strict LOSO.

All CLAHE-based representations descriptively exceed Original in pooled LOSO Accuracy and Macro-F1.

At the same time, the top-performing LOSO group is not unique: SA 30:70, Static, and Flatten are close in Accuracy, and their confidence intervals overlap substantially.

5.6. *Robustness of the Classification Conclusions*

Equivalence of the leading operating points. Table 11 and Figure 10 show that the three leading conditions are statistically equivalent at subject level within ± 0.05. SA 30:70 differs from Static CLAHE by -0.0018 Accuracy (90% CI -0.049 to 0.045; $p_{\mathrm{TOST}} = 0.047$) and $+0.0083$ Macro-F1 (-0.018 to 0.035; $p = 0.008$). Against legacy Window-wise BO the intervals are narrower still. Equivalence with jointQ is not shown: SA 30:70 is 0.070 higher in Accuracy (90% CI -0.004 to 0.143), a difference that favours the stability-aware objective but is not statistically significant. Because SA 30:70 reaches this accuracy with tSSIM 0.793 and FDV 108, against 0.721 and 311 for Static CLAHE (Table 4), the stability gain is obtained without losing more than 0.05 in subject-level Accuracy or Macro-F1.

Training-fold class weighting. With balanced class weights computed from the training participants only, pooled Neutral recall increases by 0.14–0.20 in every CLAHE condition (for example, from 0.109 to 0.300 for SA 30:70 and from 0.085 to 0.270 for Static CLAHE; Table 12). Pooled Macro-F1 rises to 0.519 (Static CLAHE), 0.515 (SA 30:70), and 0.502 (Window-wise BO). The ordering of conditions is unchanged: the same three conditions lead, jointQ and SA 70:30 trail, and Original remains lowest (Figure 11). Friedman tests are significant for Accuracy

Table 11. Paired TOST equivalence on subject-level strict-LOSO metrics ($n = 14$). Equivalence is concluded when the 90% CI lies within $\pm\delta$.

Contrast	Metric	Mean diff.	90% CI	p_{TOST} ($\delta = 0.05$)	Conclusion
SA 30:70 – Static	Accuracy	-0.0018	$[-0.049, 0.045]$	0.047	equivalent
SA 30:70 – Static	Macro-F1	0.0083	$[-0.018, 0.035]$	0.008	equivalent
SA 30:70 – Window-wise	Accuracy	-0.0052	$[-0.042, 0.031]$	0.024	equivalent
SA 30:70 – Window-wise	Macro-F1	-0.0020	$[-0.018, 0.014]$	< 0.001	equivalent
Window-wise – Static	Accuracy	0.0034	$[-0.031, 0.037]$	0.015	equivalent
Window-wise – Static	Macro-F1	0.0103	$[-0.010, 0.030]$	0.002	equivalent
SA 30:70 – jointQ	Accuracy	0.0697	$[-0.004, 0.143]$	0.679	not shown
SA 30:70 – jointQ	Macro-F1	0.0275	$[-0.022, 0.077]$	0.217	not shown

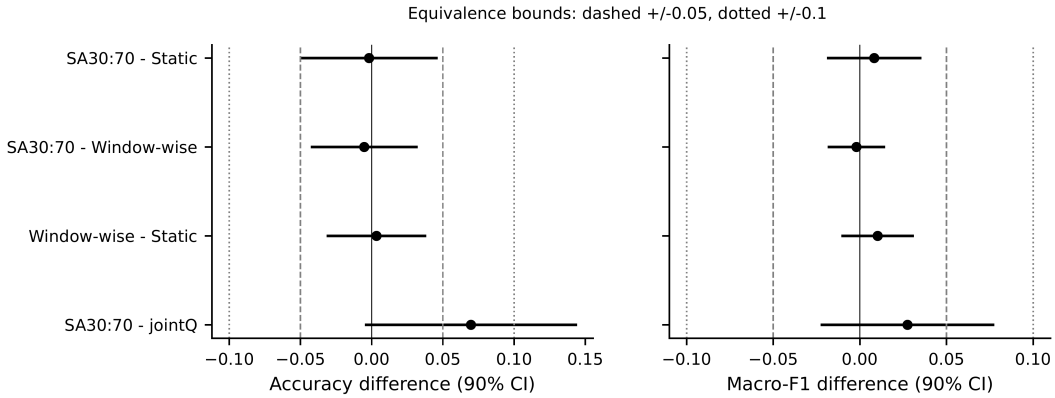


Figure 10. Paired subject-level differences with 90% confidence intervals. Dashed lines mark the primary equivalence margin (± 0.05), dotted lines the secondary margin (± 0.10).

($\chi^2 = 25.18$, $p = 0.0007$, $W = 0.257$) and Macro-F1 ($\chi^2 = 18.28$, $p = 0.011$). Unlike the unweighted experiment, five of the 56 Holm-adjusted pairwise contrasts are now significant, and all of them favour a CLAHE condition over Original (Original versus Static CLAHE: Holm $p = 0.014$ for Accuracy and Macro-F1; versus SA 30:70: $p = 0.033$ for Accuracy and $p = 0.044$ for Macro-F1; versus Window-wise BO: $p = 0.033$ for Macro-F1). Paired changes relative to the unweighted runs are positive for Macro-F1 in every condition but none survives Holm correction.

Recurrent unit. Replacing GRU(32) with LSTM(32) leaves the conclusions intact. SA 30:70 obtains the highest pooled Accuracy (0.6112) and QWK (0.3560), ahead of Static CLAHE (0.5785) and Window-wise BO (0.5734), while Original remains lowest (0.4411). This agrees with SOIC evidence on facial-landmark sequences that GRU and LSTM reach comparable accuracy [42]. The Friedman test is significant for Accuracy ($p = 0.015$) and Macro-F1 ($p = 0.026$), but no pairwise contrast survives Holm correction. jointQ shows the largest drop when moving to LSTM (Macro-F1 -0.037, raw $p = 0.035$, Holm $p = 0.35$), whereas SA 30:70 is essentially unchanged (-0.005).

Table 12. Pooled strict-LOSO results of the sensitivity classifiers. CW: GRU with training-fold class weights (all eight conditions); LSTM: LSTM(32) recurrent unit (five representative conditions).

	GRU + class weight				LSTM		
Condition	Accuracy	Macro-F1	QWK	Neutral rec.	Accuracy	Macro-F1	QWK
Original	0.4098	0.3173	-0.0098	0.1229	0.4411	0.2840	-0.0724
Static CLAHE	**0.6168**	**0.5194**	**0.3736**	0.2696	0.5785	0.4419	0.2477
Frame-wise BO	0.5710	0.4581	0.2837	0.1741	–	–	–
Window-wise BO	0.5939	0.5021	0.3236	0.2628	0.5734	0.4326	0.2579
jointQ	0.5313	0.4400	0.2329	0.2287	0.5112	0.3505	0.0723
SA 70:30	0.5271	0.4360	0.2491	0.2321	–	–	–
SA 50:50	0.5607	0.4572	0.3086	0.2287	–	–	–
SA 30:70	0.6056	0.5145	0.3471	**0.3003**	**0.6112**	**0.4495**	**0.3560**

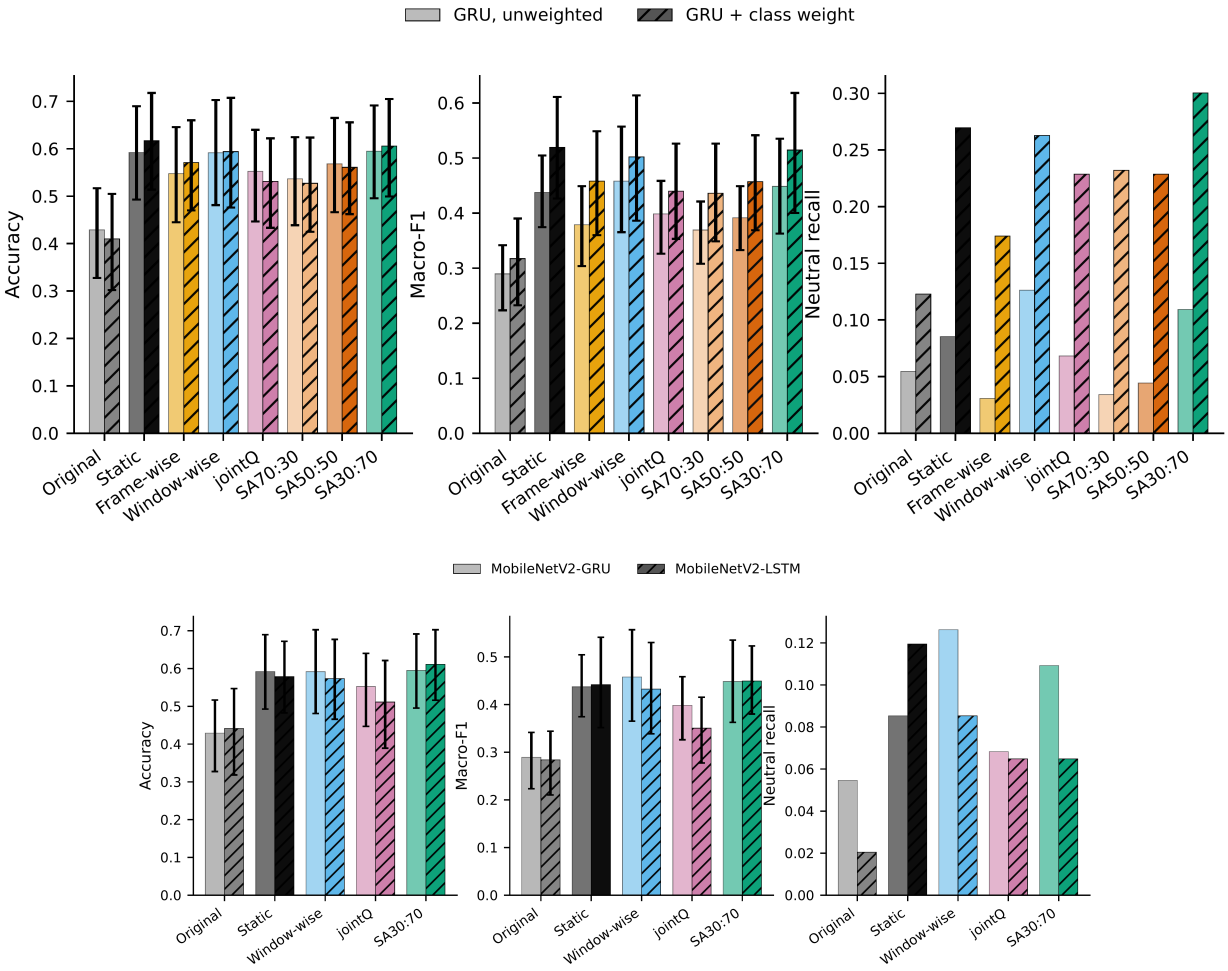


Figure 11. Sensitivity of strict-LOSO Accuracy, Macro-F1, and Neutral recall. Top: unweighted GRU (light) versus GRU with training-fold class weights (hatched). Bottom: GRU (light) versus LSTM (hatched). Error bars are 95% subject-cluster bootstrap intervals.

Window-level association between preprocessing metrics and errors. Within each representation, windows with lower FDV and higher tSSIM are classified correctly somewhat more often (Spearman $|\rho|$ between 0.07 and 0.29 for tSSIM and FDV; 41 of the 48 metric–outcome correlations, including PIQE, are significant after Holm correction). This association disappears once participants are treated as clusters: in the subject-clustered GEE with condition fixed effects, a one-standard-deviation increase in tSSIM has odds ratio 0.90 (95% CI 0.47–1.71) and in log FDV 0.83 (0.41–1.67), and normalized PIQE 1.45 (0.96–2.19). The marginal correlations therefore mostly reflect differences between participants. At condition level, Original has high tSSIM (0.786) and the lowest accuracy, so temporal stability is not sufficient on its own. Stability is best read as a property of the representation that can be raised without harming classification, not as a per-window predictor of correctness.

5.7. Optimizer Benchmark

Table 13 and Figure 12 compare search strategies under the same 13-evaluation budget. BO reaches a lower final objective than Random Search on 119 of 140 windows for jointQ and 128 of 140 for SA 30:70 (Wilcoxon, Holm $p < 0.001$; rank-biserial -0.89 and -0.92), and its final regret is about four to nine times smaller. Against the structured grid, the final objective does not differ (Holm $p = 0.36$ and 0.19), and the grid reaches low regret earlier

because its points cover the two-dimensional domain from the first evaluations. In this small search space the advantage of BO is therefore over unguided sampling rather than over a well-designed deterministic design; the BO-selected parameters reach essentially the same quality–stability operating point as the grid (tSSIM 0.793 versus 0.796 for SA 30:70).

Table 13. Budget-matched optimizer benchmark (140 windows, 13 objective calls each). Regret is measured against the best candidate found by any method; lower is better.

Objective	Method	Mean best J	Final regret	Best-known hit	Sel. tSSIM	Sel. FDV
jointQ	BO	0.3192	0.0049	67.9%	0.685	498.9
jointQ	Random Search	0.3348	0.0206	1.4%	0.694	460.8
jointQ	Structured grid	0.3187	0.0044	53.6%	0.682	508.6
SA 30:70	BO	0.2683	0.0015	81.4%	0.793	109.2
SA 30:70	Random Search	0.2802	0.0134	0.0%	0.775	149.4
SA 30:70	Structured grid	0.2684	0.0016	63.6%	0.796	100.9

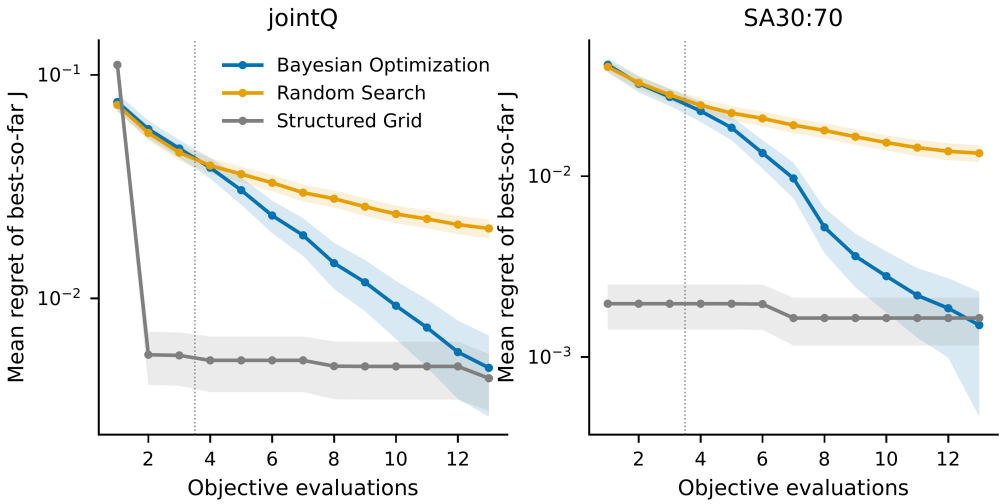


Figure 12. Mean regret of the best-so-far objective versus number of evaluations (shaded: 95% interval of the mean). The dotted line marks the end of the three initial BO evaluations.

5.8. Normalization Sensitivity

Re-selecting parameters from the evaluated candidates with per-window min–max scaling of both objective terms gives the same (c, r) pair as $Q/100$ on 89.3% of windows for SA 30:70, 91.4% for SA 70:30, and 74.3% for SA 50:50 (100% for jointQ, where scaling cannot change the ranking). The mean tSSIM of the selections changes by at most 0.017 (Table 14). Dataset-range and percentile scaling shift the balance further toward quality, because they amplify the quality term relative to the stability term; even then SA 30:70 retains higher stability than SA 50:50 under the canonical scaling. The quality–stability ordering of the weightings is therefore robust to the normalization choice, while the exact operating point of an intermediate weighting depends on it.

5.9. Preprocessing Runtime

Table 15 reports the cost of each preprocessing strategy for one ten-second, 20-frame window. Static CLAHE requires 12 ms and Original 4 ms. Every stability-aware objective requires approximately 5.6–5.9 s, the same as jointQ, so adding the temporal term to the objective has no measurable cost; the cost comes from the 13 evaluations of PIQE and SSIM. These timings are below the 10 s duration of the window on one CPU thread, whereas frame-wise BO requires 16.9 s and cannot keep pace with the input. Relative to Static CLAHE, stability-aware optimization is roughly 470 times more expensive, which is the price of the stability gain in Table 4.

Table 14. Normalization sensitivity over the evaluated candidate pool (140 windows). Agreement is the share of windows whose selected (c, r) equals the selection under $\widehat{Q} = Q/100$.

Weighting	Normalization	Agreement	Mean tSSIM	Mean FDV
SA 50:50	$Q/100$, $1 - \text{tSSIM}$ (canonical)	–	0.765	204.5
SA 50:50	Min–max per window	74.3%	0.782	145.3
SA 50:50	Min–max, dataset range	37.9%	0.704	425.0
SA 30:70	$Q/100$, $1 - \text{tSSIM}$ (canonical)	–	0.794	102.0
SA 30:70	Min–max per window	89.3%	0.796	98.5
SA 30:70	Min–max, dataset range	55.0%	0.770	184.3
SA 30:70	Robust P5–P95	66.4%	0.780	153.7

Table 15. Preprocessing runtime per 20-frame window (search plus application; 40 windows × 2 repeats, one CPU thread). Frame-wise and legacy Window-wise BO are re-implemented for timing.

Condition	Median (s)	IQR (s)	ms/frame	Relative to Static
Original	0.004	0.003–0.005	0.2	0.35×
Static CLAHE	0.012	0.012–0.013	0.6	1×
Frame-wise BO	16.93	16.47–17.31	845	1373×
Window-wise BO (legacy)	1.10	0.99–1.19	56	91×
jointQ	5.90	5.66–6.19	295	480×
SA 70:30	5.91	5.54–6.26	295	478×
SA 50:50	5.64	5.34–5.99	284	462×
SA 30:70	5.86	5.42–6.04	289	470×

5.10. Consolidated Evidence Across the Experimental Chain

Table 16 links the individual experiments to the claim each experiment is allowed to support. This synthesis is important because the evidence is layered: the temporal-granularity study identifies a representation problem, the V3 objective study demonstrates controllable stability, and strict LOSO tests whether the resulting representations remain useful for an unseen participant.

Table 16. Evidence chain from the precursor experiment to the final subject-independent conclusion.

Stage	Primary evidence	Supported conclusion
NIR method selection	Prior HE/AHE/CLAHE/Retinex comparison [8]	CLAHE is carried forward as the enhancement family; this paper does not re-claim the earlier method-selection experiment.
Temporal granularity	PIQE + within-window clip-limit drift	Independent frame optimization can improve local PIQE while changing enhancement decisions inside a sequence.
Stability-aware V3	$\widehat{Q}$, tSSIM, FDV, Friedman, Spearman	Objective weighting systematically controls the quality–stability operating point.
Strict LOSO	pooled metrics + cluster-bootstrap CI + subject blocks	CLAHE-based representations remain consequential when the test participant is unseen.
Inference boundary	Friedman + Wilcoxon-Holm	No individual preprocessing pair is established as universally superior after multiplicity correction.
Equivalence	Paired TOST, $\delta = 0.05$	SA 30:70, Static CLAHE, and Window-wise BO are equivalent in subject-level Accuracy and Macro-F1; SA 30:70 is far more stable.
Robustness	Class-weighted GRU, LSTM	The ordering of conditions holds; class weighting improves Neutral recall and makes CLAHE significantly better than Original.
Search and cost	Optimizer benchmark, runtime	BO beats Random Search and matches a structured grid; the stability term adds no cost beyond jointQ.

6. Discussion

6.1. What Strict LOSO Changes

The strict LOSO experiment substantially changes the interpretation relative to the window-level protocol.

The window-level split answers whether different preprocessing representations can be learned under a partition in which participant identity is not explicitly separated.

Strict LOSO instead asks whether the same preprocessing design transfers to a participant whose facial appearance is completely absent from training.

The decrease from approximately 0.88–0.95 Accuracy under the window-level protocol to approximately 0.43–0.59 under LOSO indicates that the former estimates were optimistic for the unseen-subject use case.

We therefore do not describe the earlier protocol as deterministic "leakage." More precisely, the protocol did not enforce the participant independence required for an unseen-subject generalization claim.

The LOSO result also strengthens the apparent relevance of preprocessing. Importantly, this does not erase the precursor temporal-granularity result. The earlier PIQE/drift experiment and the final LOSO experiment answer different levels of the same question: the first characterizes what the preprocessor does to a sequence, whereas the second characterizes whether a representation learned from 13 participants transfers to the fourteenth. Reading these stages together prevents a high window-level classifier score from being mistaken for evidence of unseen-subject generalization.

Original falls to 0.4290 pooled Accuracy, whereas every CLAHE-based representation exceeds 0.536.

This does not establish pairwise statistical superiority of any individual CLAHE strategy after multiplicity correction. It does, however, show that representation design remains consequential when participant-specific facial appearance cannot be shared between training and test sets.

6.2. Temporal Stability Is a Design Dimension, Not a Universal Surrogate

At the representation level, increasing stability weight clearly raises tSSIM and lowers FDV.

Under LOSO, the same ordering is accompanied by improvement across the three stability-aware classifier metrics from SA 70:30 to SA 30:70, although jointQ lies slightly above SA 70:30, so the response over the full weight range is not monotonic.

Accordingly, the earlier mechanistic interpretation attributing the non-monotonic response to a specific balance between enhancement strength and nuisance variation is not retained as a principal conclusion.

The more defensible conclusion is narrower: temporal stability is a consequential preprocessing design dimension under subject-independent sequence learning, but tSSIM and FDV remain proxies for visual temporal consistency rather than direct measures of feature-space stability or drowsiness semantics.

The present experiment does not establish a causal chain from CLAHE parameter changes to embedding displacement and subsequent GRU error. The subject-clustered window-level analysis points the same way: once participant differences are accounted for, tSSIM and FDV do not predict whether an individual window is classified correctly.

What the evidence does support is a dominance argument. SA 30:70 is equivalent to the strongest baselines in subject-level classification, is descriptively about 0.07 higher in Accuracy than the quality-only objective, and produces windows with markedly higher structural consistency and lower frame-difference variance. The same ordering of conditions is observed with the GRU, class-weighted GRU, and LSTM classifiers. When two representations classify equally well, the more stable one is preferable for any downstream use that is sensitive to spurious inter-frame change, such as alarm smoothing or temporal feature extraction.

6.3. No Single Universal Winner

SA 30:70 has the highest pooled Accuracy and QWK.

Flatten/legacy Window-wise BO has the highest pooled Macro-F1 and the highest Neutral recall.

Static CLAHE is nearly tied in Accuracy while requiring no per-window parameter search.

Because no pairwise contrast survives Holm correction and the three leading conditions are formally equivalent within ± 0.05, the data do not support a universal superiority claim.

These three conditions are better interpreted as different operating points.

For a practitioner whose priority is implementation simplicity and strong unseen-subject performance, Static CLAHE provides a credible reference configuration.

When explicit temporal consistency is a primary preprocessing objective and per-window optimization is acceptable, SA 30:70 provides the strongest descriptive Accuracy and QWK among the tested stability-aware objectives, at about 5.9 s of single-thread CPU time per ten-second window.

When the Neutral state matters, training-fold class weighting should be combined with any of these representations, because it raises Neutral recall from roughly 0.1 to 0.26–0.30 without reducing Accuracy.

When balanced class discrimination is emphasized, Flatten/legacy Window-wise BO provides the highest pooled Macro-F1 and Neutral recall.

These guidelines are conditional on the present dataset, classifier, and experimental setting and should not be interpreted as deployment prescriptions.

6.4. Optimizer Choice, Normalization, and Deployment Scope

BO is retained because the per-window objective is treated as an expensive black-box function and the study uses a limited 13-evaluation budget [12,13].

The benchmark shows that this choice is clearly better than random sampling but not better than a carefully placed 13-point grid in the present two-parameter space. A practitioner who keeps the same domain can therefore use a fixed grid with no loss of quality; BO becomes more useful when the domain is enlarged, for example with additional enhancement parameters, where a grid of equal size covers the space far more sparsely. Genetic algorithms were not benchmarked.

The $Q/100$ normalization is an operational scaling supported by the empirical PIQE range of this dataset; per-window min–max scaling selects the same parameters for most windows, but a dataset or metric with a different range would need recalibration.

Stability-aware preprocessing costs about 5.9 s per ten-second window on one CPU thread, and the classifier adds approximately 255–274 ms. The pipeline thus keeps pace with a ten-second decision interval in this unoptimized Python implementation, but not with frame-rate updates. Parallel evaluation of the 13 candidates, a cheaper quality surrogate, or a fixed grid would reduce this cost.

We do not claim real-time readiness, false-alarm performance, or time-to-alarm feasibility.

Controlled end-to-end latency, memory, energy, and hardware-specific profiling remain deployment-oriented tasks rather than evidence supplied by the present experiment.

7. Limitations

First, the study uses one NIR corpus with 14 participants and 36 usable sessions.

Strict LOSO provides subject-independent validation within DROZY, but it does not establish cross-dataset, cross-camera, or cross-population generalization.

Second, all eight representations are evaluated with one primary classifier, MobileNetV2–GRU, with an LSTM variant for five conditions as a sensitivity check. Both share the MobileNetV2 backbone, so robustness to temporal convolutional or transformer-based classifiers remains untested.

Third, the subject mapping used for LOSO was reconstructed and verified from the original data-generation traversal because the saved feature arrays did not originally contain subject identifiers.

Although every window received a verified mapping and the pooled test accounting is exact, preserving the subject/session manifest during the original data-generation stage would have been preferable.

Fourth, tSSIM and FDV are visual temporal-consistency proxies.

No optical-flow, landmark-stability, or embedding-space analysis is included; therefore causal claims concerning representation dynamics are avoided.

Fifth, class imbalance remains substantial and the Neutral class remains particularly difficult.

Training-fold class weighting was evaluated as a sensitivity analysis and improved Neutral recall, but focal loss, resampling, and ordinal-regression objectives were not tested.

QWK is therefore used as an ordinal evaluation metric rather than as evidence that the classifier was trained using an ordinal objective.

Sixth, the scalarization weights and $Q/100$ scaling are operational design choices evaluated through a limited sensitivity set rather than through a calibration-free multi-objective procedure.

Seventh, legacy Frame-wise and Window-wise BO use a narrower clip domain with fixed 8×8 tiles, whereas the V3 objectives search $c \in [0.1, 8.0]$ and $r \in \{4, \ldots, 8\}$. Comparisons between legacy and V3 conditions therefore combine differences in granularity, search domain, and objective; the comparisons among the four V3 objectives do not.

Finally, the optimizer benchmark covers BO, Random Search, and a structured grid but not genetic algorithms, runtime was measured for preprocessing on one CPU thread rather than end-to-end on embedded hardware, and power consumption, false-alarm rate, and time-to-alarm were not measured.

The present evidence therefore supports a preprocessing design formulation and subject-independent within-dataset validation, rather than claims of optimizer supremacy or deployment readiness.

Data and Code Availability

The experimental configuration, protocols, and results needed to interpret and reproduce the reported analyses are documented in this manuscript. The DROZY videos are available from their maintainers [1]. Supporting artifacts, including the verified subject/window manifest, strict-LOSO predictions, statistical tables, analysis notebooks, derived face arrays, and trained model checkpoints, are available from the corresponding author on reasonable request, where available.

8. Conclusion

This study reframes CLAHE optimization for NIR drowsiness video around the temporal unit consumed by a sequence classifier.

Frame-wise BO achieves strong harmonized PIQE but introduces within-window parameter variation, motivating a video-window-wise formulation in which one parameter pair is applied consistently across the complete 20-frame input.

The stability-aware objective then makes the quality–temporal-consistency trade-off explicit.

The primary downstream validation uses strict subject-independent LOSO.

Across 112 LOSO training runs, SA 30:70 attains the highest pooled Accuracy and QWK, Flatten/legacy Window-wise BO attains the highest Macro-F1, and Static CLAHE remains highly competitive.

Among the stability-weighted objectives, a larger temporal weight is accompanied by better pooled LOSO metrics, and SA 30:70 is descriptively higher than the quality-only objective.

No pairwise subject-level comparison among the leading conditions is significant after Holm correction; instead, equivalence testing shows that SA 30:70, Static CLAHE, and legacy Window-wise BO classify unseen participants equally well within ± 0.05, while SA 30:70 produces far more temporally stable input. These conclusions hold with class-weighted training, which also makes the CLAHE conditions significantly better than unenhanced input, and with an LSTM classifier. The stability term adds no computational cost beyond quality-only optimization, BO outperforms random search and matches a structured grid, and the results are insensitive to min–max normalization of the objective.

The principal contribution is therefore not a new CLAHE algorithm, a new neural architecture, or a universally optimal parameter setting.

It is the formulation and empirical evaluation of temporal granularity and explicit temporal stability as preprocessing design variables for subject-independent sequence learning.

Future work should test this formulation across datasets and non-recurrent temporal architectures, extend the search space where model-based optimization is expected to pay off, evaluate ordinal-aware losses, and perform end-to-end deployment profiling on embedded hardware.

Acknowledgement

The authors acknowledge Universitas Dian Nuswantoro and Universitas Islam Negeri Walisongo for supporting this research and the maintainers of the DROZY database for making the dataset available to the research community.

REFERENCES

1. Q. Massoz, T. Langohr, C. Francois, and J. G. Verly, *The ULg multimodality drowsiness database (called DROZY) and examples of use*, in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–7, 2016.
2. A. Fawzi, A. Achuthan, and B. Belaton, *Adaptive Clip Limit Tile Size Histogram Equalization for Non-Homogenized Intensity Images*, IEEE Access, vol. 9, pp. 164466–164492, 2021.
3. M. B. Mortatha, S. S. Thabit, H. R. Abdul Ameer, and R. R. Nuiaa, *Dynamic Clip Limit Window Size Histogram Equalization for Poor Information Images*, International Journal of Intelligent Engineering and Systems, vol. 15, no. 5, pp. 57–70, 2022.
4. U. Kuran and E. C. Kuran, *Parameter selection for CLAHE using multi-objective cuckoo search algorithm for image contrast enhancement*, Intelligent Systems with Applications, vol. 12, 200051, 2021.
5. M. Braik, M. A. Al-Betar, M. A. Mahdi, M. Al-Shalabi, S. Ahamad, and S. A. Saad, *Enhancement of satellite images based on CLAHE and augmented elk herd optimizer*, Artificial Intelligence Review, vol. 58, no. 2, 38, 2024.
6. F. Alzami, M. Naufal, H. Al Azies, S. Winarno, and M. A. Soeleman, *Time Distributed MobileNetV2 with Auto-CLAHE for Eye Region Drowsiness Detection in Low Light Conditions*, International Journal of Advanced Computer Science and Applications, vol. 15, no. 11, 2024.
7. F. Alzami, M. Naufal, R. S. Basuki, S. Winarno, H. Al Azies, S. L. Lutfi, and R. M. Brilianto, *Bayesian-Optimized CLAHE for Enhanced Drowsiness Detection in Low-Light Conditions Using Time-Distributed MobileNetV2-GRU Architecture*, Statistics, Optimization & Information Computing, vol. 15, no. 1, pp. 274–294, 2025.
8. M. H. Subowo, P. N. Andono, G. F. Shidik, and H. A. Santoso, *Comparison of Histogram Equalization and Multi-Scale Retinex Methods for Near-Infrared Image Enhancement in Drowsiness Detection*, International Journal of Advanced Computer Science and Applications, vol. 17, no. 1, pp. 846–858, 2026.
9. M. H. Subowo, P. N. Andono, G. F. Shidik, and H. A. Santoso, *A Decade of Progress in Drowsiness Detection Using Facial Features: A Comprehensive Bibliometric Analysis and Thematic Mapping*, Statistics, Optimization & Information Computing, vol. 15, no. 5, pp. 3572–3595, 2026.
10. N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani, *Blind image quality evaluation using perception based features*, in 2015 Twenty First National Conference on Communications (NCC), pp. 1–6, IEEE, 2015.
11. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, *Image quality assessment: from error visibility to structural similarity*, IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.
12. D. R. Jones, M. Schonlau, and W. J. Welch, *Efficient Global Optimization of Expensive Black-Box Functions*, Journal of Global Optimization, vol. 13, pp. 455–492, 1998.
13. J. Snoek, H. Larochelle, and R. P. Adams, *Practical Bayesian Optimization of Machine Learning Algorithms*, Advances in Neural Information Processing Systems, vol. 25, 2012.
14. M. Sandler, A. Howard, M. Zhu, M. Zhmoginov, and L.-C. Chen, *MobileNetV2: Inverted Residuals and Linear Bottlenecks*, in Proceedings of CVPR, pp. 4510–4520, 2018.
15. K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*, in Proceedings of EMNLP, pp. 1724–1734, 2014.
16. A. Quddus, A. S. Zandi, L. Prest, and F. J. E. Comeau, *Using long short term memory and convolutional neural networks for driver drowsiness detection*, Accident Analysis & Prevention, vol. 156, 106107, 2021.
17. S. E. Bekhouche, Y. Ruichek, and F. Dornaika, *Driver drowsiness detection in video sequences using hybrid selection of deep features*, Knowledge-Based Systems, vol. 252, 109436, 2022.
18. J. Bai, W. Yu, Z. Xiao, V. Havyarimana, A. C. Regan, H. Jiang, and L. Jiao, *Two-Stream Spatial–Temporal Graph Convolutional Networks for Driver Drowsiness Detection*, IEEE Transactions on Cybernetics, vol. 52, no. 12, pp. 13821–13833, 2022.
19. S. Fa, X. Yang, S. Han, Z. Feng, and Y. Chen, *Multi-scale spatial–temporal attention graph convolutional networks for driver fatigue detection*, Journal of Visual Communication and Image Representation, vol. 93, 103826, 2023.
20. L. Yang, H. Yang, H. Wei, Z. Hu, and C. Lv, *Video-Based Driver Drowsiness Detection With Optimised Utilization of Key Facial Features*, IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 7, pp. 6938–6950, 2024.
21. N. Bonneel, J. Tompkin, K. Sunkavalli, D. Sun, S. Paris, and H. Pfister, *Blind Video Temporal Consistency*, ACM Transactions on Graphics, vol. 34, no. 6, Article 196, 2015.
22. W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.-H. Yang, *Learning Blind Video Temporal Consistency*, in Proceedings of the European Conference on Computer Vision (ECCV), pp. 170–185, 2018.
23. M. Ahmed, S. Masood, M. Ahmad, and A. A. Abd El-Latif, *Intelligent Driver Drowsiness Detection for Traffic Safety Based on Multi CNN Deep Model and Facial Subsampling*, IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 10, pp. 19743–19752, 2022.

24. A.-C. Phan, N.-H.-Q. Nguyen, T.-N. Trieu, and T.-C. Phan, *An Efficient Approach for Detecting Driver Drowsiness Based on Deep Learning*, Applied Sciences, vol. 11, no. 18, 8441, 2021.
25. M.-Z. Liu, X. Xu, J. Hu, and Q.-N. Jiang, *Real time detection of driver fatigue based on CNN-LSTM*, IET Image Processing, vol. 16, no. 2, pp. 576–595, 2022.
26. A.-C. Phan, T.-N. Trieu, and T.-C. Phan, *Driver drowsiness detection and smart alerting using deep learning and IoT*, Internet of Things, vol. 22, 100705, 2023.
27. S. P. Soman, G. S. Kumar, S. B. Nuthalapati, S. Zafar, and A. K. M., *Internet of things assisted deep learning enabled driver drowsiness monitoring and alert system using CNN-LSTM framework*, Engineering Research Express, vol. 6, no. 4, 045239, 2024.
28. M. Kim and G. Choi, *STFTransNet: A Transformer Based Spatial Temporal Fusion Network for Enhanced Multimodal Driver Inattention State Recognition System*, Sensors, vol. 25, no. 18, 5819, 2025.
29. E. Civik and U. Yuzgec, *Real-time driver fatigue detection system with deep learning on a low-cost embedded system*, Microprocessors and Microsystems, vol. 99, 104851, 2023.
30. N. N. Alajlan and D. M. Ibrahim, *DDD TinyML: A TinyML-Based Driver Drowsiness Detection Model Using Deep Learning*, Sensors, vol. 23, no. 12, 5696, 2023.
31. S. Essahraui, I. Lamaakal, I. El Hamly, Y. Maleh, I. Ouahbi, K. El Makkaoui, M. F. Bouami, P. Plawiak, O. Alfarraj, and A. A. Abd El-Latif, *Real-Time Driver Drowsiness Detection Using Facial Analysis and Machine Learning Techniques*, Sensors, vol. 25, no. 3, 812, 2025.
32. S. Cao, P. Feng, W. Kang, Z. Chen, and B. Wang, *Optimized driver fatigue detection method using multimodal neural networks*, Scientific Reports, vol. 15, no. 1, 12240, 2025.
33. J. Bergstra and Y. Bengio, *Random Search for Hyper-Parameter Optimization*, Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012.
34. S. Holm, *A Simple Sequentially Rejective Multiple Test Procedure*, Scandinavian Journal of Statistics, vol. 6, no. 2, pp. 65–70, 1979.
35. D. J. Schuirmann, *A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability*, Journal of Pharmacokinetics and Biopharmaceutics, vol. 15, no. 6, pp. 657–680, 1987.
36. K.-Y. Liang and S. L. Zeger, *Longitudinal Data Analysis Using Generalized Linear Models*, Biometrika, vol. 73, no. 1, pp. 13–22, 1986.
37. X. Wang and A. Gupta, *Unsupervised Learning of Visual Representations Using Videos*, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2794–2802, 2015.
38. I. Misra, C. L. Zitnick, and M. Hebert, *Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification*, in Proceedings of the European Conference on Computer Vision (ECCV), pp. 527–544, 2016.
39. R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui, *Spatiotemporal Contrastive Video Representation Learning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6964–6974, 2021.
40. G. Eilertsen, R. K. Mantiuk, and J. Unger, *Single-Frame Regularization for Temporally Stable CNNs*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11176–11185, 2019.
41. H. A. Kazum, A. M. Al-Bakry, and J. Waleed, *Next-Generation Driver Monitoring Systems: A Review of Multimodal Detection of Distraction, Drowsiness, and Sudden Medical Conditions*, Statistics, Optimization & Information Computing, vol. 15, no. 6, pp. 4553–4576, 2026.
42. S. Winarno, F. Alzami, D. A. Santoso, M. Naufal, H. Al Azies, R. M. Brilianto, and K. A/P Sonai Muthu, *Efficient GRU-based Facial Expression Recognition with Adaptive Loss Selection*, Statistics, Optimization & Information Computing, vol. 14, no. 6, pp. 3468–3499, 2025.