STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, December 2026, pp 5247–5270.
Published online in International Academic Press (www.IAPress.org)

Top-K Landmark Selection, Optical-Flow Fusion, and BorderlineSMOTE for Subject-Independent Facial Expression Recognition: A Leakage-Aware Cross-Dataset Evaluation

Sri Winarno[1,*], Farrikh Alzami[1], Dewi Agustini Santoso[1], Muhammad Naufal[1], Harun Al Azies[1], Kalaiarasi A/P Sonai Muthu[2]

[1] *Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, 50131, Indonesia*
[2] *Faculty of Information Science and Technology, Multimedia University, Melaka, 75450, Malaysia*

Abstract Facial expression recognition (FER) systems increasingly rely on landmark-based and motion-based representations to reduce sensitivity to illumination and identity, yet no published pipeline jointly combines Top-K landmark selection, optical-flow motion fusion, and resampling-based class-imbalance handling under a subject-disjoint evaluation protocol. This study proposes a CNN-GRU architecture that selects informative facial landmarks using an optical-flow-derived importance score, fuses geometric position and motion features, and applies BorderlineSMOTE oversampling to counter class imbalance, evaluated under a leakage-free, subject-disjoint cross-validation protocol on two datasets of contrasting scale: the Extended Cohn-Kanade dataset (CK+, 106 subjects) and the Indonesian Mixed Emotion Dataset (IMED, 15 subjects), each restricted to six basic emotion classes. Across five seeds, ten subject-disjoint folds, five landmark retention levels, three input modalities, and two resampling conditions, Friedman and Wilcoxon signed-rank tests with 95% confidence intervals show that landmark retention of 20 to 30 percent and BorderlineSMOTE significantly improve weighted F1 on CK+ (F1w = 0.810) and, more weakly, on IMED (F1w = 0.426), while modality fusion significantly outperforms single-modality input on CK+ but not on IMED. On IMED, the BorderlineSMOTE result is the one finding whose significance depends on the multiple-comparison correction applied: it survives Benjamini-Hochberg FDR correction but not the stricter Bonferroni correction, and is reported with this dependency stated explicitly rather than under a single, correction-agnostic conclusion. A Dataset-by-Factor interaction test shows all three effects, not only modality, are significantly smaller on IMED than on CK+; modality attenuates most severely, to roughly one-eighth of its CK+ size, which is why it alone loses significance within IMED, whereas landmark pruning and resampling retain enough of their larger CK+ margin to stay significant despite comparable proportional shrinkage. Every effect weakens on the smaller dataset, but only the narrowest-margin effect loses significance outright. This specific pattern has not previously been isolated under a controlled, identical-protocol comparison. Given IMED's fifteen-subject pool, its findings are reported as exploratory and dataset-limited rather than confirmatory.

Keywords facial expression recognition; landmark selection; optical flow; class imbalance; BorderlineSMOTE; CNN-GRU; subject-independent evaluation

DOI: 10.19139/soic-2310-5070-4519

1. Introduction

Facial expression recognition (FER) infers a person's emotional state from facial visual cues, underlying applications in human-computer interaction, driver monitoring, education analytics, and clinical affect assessment. This study targets the six basic emotion categories shared across the FER literature and by the two datasets evaluated here: anger, disgust, fear, happiness, sadness, and surprise. FER remains difficult because expressions

*Correspondence to: Sri Winarno (Email: sri.winarno@dsn.dinus.ac.id). Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, 50131, Indonesia.

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

unfold over short temporal windows, are confounded by identity-specific facial geometry, and because class distributions in naturally elicited datasets are rarely balanced.

Two feature families dominate FER: appearance-based methods operating on pixel intensities via CNNs, and landmark-based methods tracking a sparse set of fiducial points, which are more robust to illumination and identity but depend on landmark-detector accuracy. A related family adds a temporal or motion component, most often optical flow, to capture displacement rather than only position. This study also uses the term *optical flow* for its motion channel; precisely, this channel is a finite-difference displacement of tracked landmark coordinates between frames, not a dense, pixel-intensity-based flow field (Section 3.1 states this explicitly). "Optical flow" is retained as the running name throughout for continuity with the terminology the pipeline's other components (Top-K selection, modality fusion) are described against. For sequence classification, CNN-GRU and CNN-LSTM hybrids are now common; GRU uses fewer gates than LSTM and converges faster while matching its accuracy [1]. Recent CNN-GRU pipelines report strong CK+ results: Res-RBG reports 98.10% [2], and a cascade-attention ResNeXt-GRU network reports 99.23% [3].

Three gaps motivate this study. First, no published pipeline jointly combines Top-K landmark selection, optical-flow fusion, and resampling-based imbalance handling within a CNN-GRU architecture; landmark-centric methods [4, 5] use landmarks as a static or alignment feature without a data-driven pruning mechanism, and landmark-flow fusion has been demonstrated only for micro-expression recognition [6]. Second, class-imbalance handling in FER is dominated by cost-sensitive losses [7–9] rather than resampling; BorderlineSMOTE [10], which extends SMOTE [11] by oversampling only boundary-adjacent minority examples, remains comparatively unexplored in FER. Third, subject-independent evaluation is inconsistently applied in macro-FER even though non-subject-disjoint splits are known to inflate reported accuracy elsewhere [12, 13]; within FER, Li and Deng [14] show within-dataset CK+ accuracy above 90% collapsing to roughly 50–60% under cross-dataset evaluation. The Indonesian Mixed Emotion Dataset (IMED) [15] adds a further dimension: its small, 15-subject pool makes subject-disjoint evaluation especially demanding, and no prior study evaluates Top-K selection, optical-flow fusion, or BorderlineSMOTE on IMED under such a protocol.

This study extends a preliminary investigation that compared GRU and LSTM under a one-vs-all framework with adaptive loss selection on CK+ alone, using a single evaluation condition per configuration ($n = 6$) [16]. The present study is a substantially larger successor: a full multi-class CNN-GRU classifier, an explicit resampling ablation in place of loss adaptation, Top-K selection and optical-flow fusion as additional factors, two datasets instead of one, and a five-seed, ten-fold, subject-disjoint design suitable for Friedman, Nemenyi, and Wilcoxon testing.

These gaps motivate three research questions. RQ1: does combining Top-K landmark selection with optical-flow fusion improve subject-independent FER over single-modality representations? RQ2: does BorderlineSMOTE improve subject-independent FER relative to no resampling? RQ3: do the RQ1/RQ2 effects observed on a large-subject-pool dataset generalize to a much smaller one evaluated under an identical protocol? Correspondingly, RO1 quantifies the effect of five landmark retention levels (20, 30, 40, 50, 100%) and three modalities (geometric, optical-flow, combined) on FER performance; RO2 quantifies the effect of BorderlineSMOTE versus no resampling; RO3 compares the RO1/RO2 effects across CK+ (106 subjects) and IMED (15 subjects) under an identical, leakage-free protocol.

The primary contribution of this study is methodological integration rather than any single new component: Top-K landmark selection, optical-flow (landmark-displacement) fusion, and BorderlineSMOTE are each individually established, and the contribution is combining them inside one CNN-GRU pipeline under a single, leakage-free, subject-disjoint protocol applied identically to two datasets of very different scale, with a formal statistical framework that tests whether the resulting effects transfer between them. Specifically, this study contributes: (1) a CNN-GRU pipeline combining optical-flow-guided Top-K landmark selection with explicit modality fusion, a combination not previously evaluated together in macro-expression FER; (2) BorderlineSMOTE evaluated as a controlled, isolated ablation against a no-resampling baseline, addressing a resampling-versus-loss-reweighting gap in FER; (3) an identical, leakage-free, subject-disjoint protocol applied to two datasets differing by an order of magnitude in subject count; (4) Friedman, Nemenyi, and Wilcoxon testing with 95% confidence intervals over fold-level means, reported separately per dataset; and (5) a formal Dataset-by-Factor interaction test showing all three

effects, not only modality, are significantly attenuated on the smaller dataset, isolating why modality alone loses significance there. Novelty is claimed at the level of this combined evaluation and interaction-testing framework, not at the level of any individual component.

2. Related Work

2.1. Landmark Selection, Optical-Flow Fusion, and Temporal Modeling

Landmark-centric FER methods typically use landmarks in full or grouped by anatomical region rather than pruned by a data-driven score. Liu et.al fuse spatial-attention CNN features with attention-LSTM features over seven facial regions (95.15% CK+, 98.57% JAFFE, image-level ten-fold split) [4]. Ayeche and Alti classify a Delaunay triangulation of AAM landmarks with classical classifiers, reporting 100% accuracy on CK+ under an 80/20 split with no subject grouping stated (a separate 96.94% figure in the same paper is for KDEF, not CK+) [17]. Soylemez and Ergen restrict CNN training to a landmark-derived region of interest, raising CK+ accuracy from 84% to 98.04%; uniquely among the CK+ studies reviewed here, this study explicitly enforces subject-disjoint 8-fold splits, though its task is single-frame rather than sequence classification [18]. Graph-based representations [5, 19, 20] likewise treat landmark topology as a static, single-frame feature without a motion channel.

Optical flow as a primary descriptor is comparatively well established only in micro-expression recognition: AM3F-FlowNet fuses horizontal/vertical flow and strain with attention and a logit-adjusted loss [21]; ARTNet amplifies motion before computing flow [22]; and a dual-stream network fusing a landmark branch with an optical-flow branch [6] is the closest methodological analog to this study, but confined to micro-expression benchmarks. Within macro-FER, ERFN fuses Farneback optical flow with VGG16/InceptionResNetV2 appearance features, reporting 98.00% on a six-class, 309-sequence CK+ subset under an image-level 50/25/25 split [23], but uses appearance rather than landmark features. No macro-FER pipeline combines Top-K landmark selection with optical-flow fusion inside a CNN-GRU architecture, which is the gap RO1 addresses.

2.2. Class-Imbalance Handling

Cost-sensitive losses dominate recent FER imbalance handling: focal loss [7] underlies AM3F-FlowNet and M3ENet [24]; dynamic intra-class clustering [8] and attention-based affinity sharing (RASN) [9] address imbalance architecturally; weighted cross-entropy and dynamic augmentation are also used [25, 26]. Resampling alternatives are comparatively rare: BorderlineSMOTE [10] extends SMOTE [11] by oversampling only boundary-adjacent minority examples; DeepSMOTE performs SMOTE interpolation in a learned latent space [27]; outside FER, a mixed SMOTE-CNN model reports 99.08% averaged across 24 imbalanced benchmarks [28], and a comparison of oversampling variants finds SMOTE and ADASYN most consistently competitive [29]. A cautionary finding directly relevant to this study's design is that oversampling applied before, rather than within, a cross-validation split can produce overly optimistic results [30], which motivates the fold-internal SMOTE application described in Section 3.5. Borderline-SMOTE specifically remains rare as a resampling choice for facial-expression data, motivating RO2.

2.3. Subject-Independent Evaluation and Cross-Dataset Generalization

Non-subject-disjoint splits allow identity-specific geometry to leak into test performance. For example there exists document leakage across 294 papers in 17 fields [13]; Yagis et.al show slice-level rather than subject-level cross-validation inflated brain-MRI accuracy by 29–55 percentage points across four datasets [12]; Rosenblat et.al show feature-selection and repeated-subject leakage produce the largest inflation among five leakage types tested, worsening as sample size shrinks [31]. Within FER, Kas et.al report a rare rigorous example, a person-independent, leave-one-subject-out framework across five macro-FER datasets [32]. Li and Deng quantify FER dataset bias directly: within-dataset accuracy above 90% on CK+, JAFFE, MMI, and Oulu-CASIA collapses to roughly 50–60% cross-dataset [14]. Several CK+ results already discussed, together with an improved VGG19 network (93.94%) [33] and a ResNet50-GRU hybrid (95.56%, seven-class) [34], do not state whether their evaluation

is subject-disjoint; Bi et.al note directly that CK+ has no designated train/validation/test split, so evaluation methodology varies across studies that use it. Two prior studies evaluate IMED [33]. Mahastama et.al train a stacked CNN jointly on FER2013, CK+, and IMED, reporting 71% combined accuracy without isolating IMED-specific, subject-disjoint performance [35]. Akeh and Kusuma fine-tune three architectures (Vision Transformer, Swin Transformer v2, ConvNeXt) on an eighteen-category label set spanning all of IMED's twelve mixed-emotion and six basic-emotion categories jointly, rather than the six-class basic-emotion subset used here, reporting 79.37% test accuracy with their best model (Vision Transformer) [36]; this is not directly comparable to a six-class figure. Neither study reports a subject-disjoint protocol: Mahastama et al.'s is not stated, while Akeh and Kusuma's own description is a randomly shuffled, class-stratified split over individual images (60/20/20) with no stated subject-level grouping, so, given IMED's fifteen-subject, 281-video structure, same-subject frames plausibly span their training, validation, and test sets. No prior study applies an identical, leakage-free, subject-disjoint protocol to two FER datasets differing by an order of magnitude in subject count, which is the gap RO3 addresses.

3. Methodology

This section describes the proposed pipeline: landmark and motion features (3.1), Top-K landmark selection (3.2), input modalities (3.3), the CNN-GRU architecture (3.4), BorderlineSMOTE handling (3.5), and the leakage-free training and evaluation procedure (3.6). Figure 1 summarizes the pipeline; dataset-specific quantities and hyperparameter values are given in Section 4.

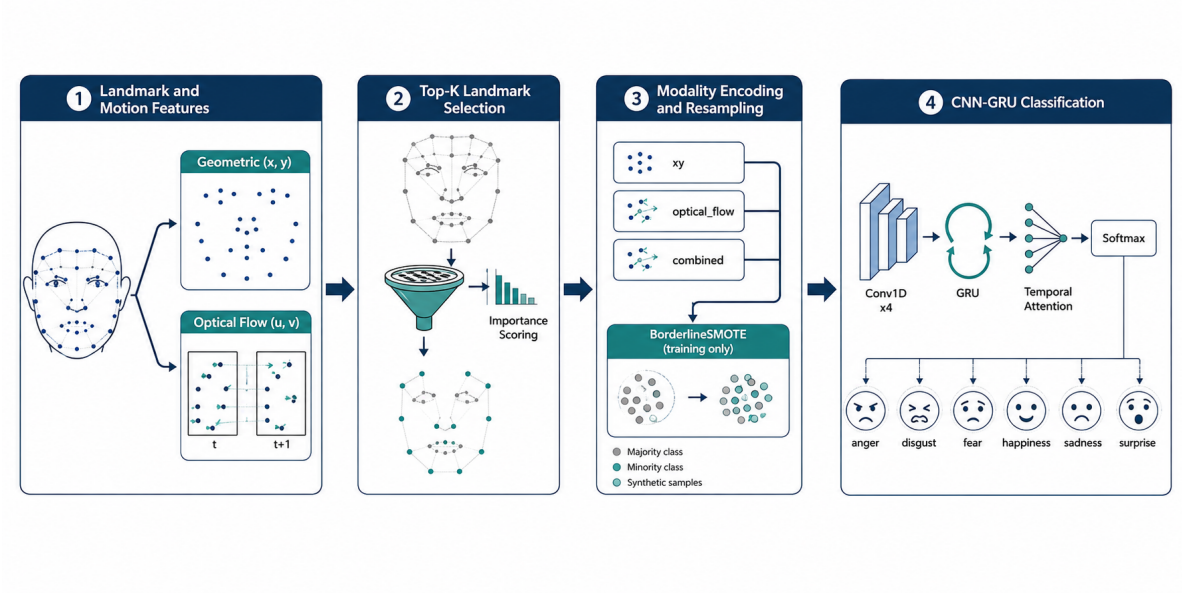


Figure 1. Overview of the proposed pipeline.

Raw landmark sequences are converted into geometric and optical-flow branches; the flow branch drives Top-K landmark selection; selected landmarks are encoded under one of three modalities; class imbalance in the resulting inner-training tensor is optionally corrected with BorderlineSMOTE; and a CNN-GRU network with temporal attention produces the prediction. Algorithm 1 (Section 3.6) states the full procedure formally.

3.1. Landmark and Motion Feature Representation

Each sample is a sequence of T frames, each frame represented by L facial landmarks located with a 468-point face mesh detector (MediaPipe Face Mesh), so $L = 468$ for both datasets. Let $x_{i,t}, y_{i,t}$ denote the position of landmark

$i \in \{1, \ldots, L\}$ at frame $t \in \{1, \ldots, T\}$, normalized to the unit interval. A motion feature is computed as the frame-to-frame displacement of each landmark, a finite-difference velocity rather than a dense, image-intensity-based optical-flow field:

$$u_{i,t} = x_{i,t+1} - x_{i,t}, \qquad v_{i,t} = y_{i,t+1} - y_{i,t}, \qquad t \in \{1, \ldots, T-1\}. \tag{1}$$

Following the terminology used throughout this paper, $(u_{i,t}, v_{i,t})$ is referred to as the optical-flow feature, with Equation 1 making explicit its derivation from tracked coordinates rather than a Lucas-Kanade- or Farneback-style algorithm. This yields $T' = T - 1$ motion frames; the geometric branch is correspondingly re-indexed to $t \in \{2, \ldots, T\}$.

3.2. Optical-Flow-Guided Top-K Landmark Selection

Not every landmark is equally informative; landmarks on the outer contour or forehead exhibit little expression-related motion. For N training samples, the flow magnitude of landmark i at frame t in sample n is $m_{i,t}^{(n)} = \sqrt{(u_{i,t}^{(n)})^2 + (v_{i,t}^{(n)})^2}$, and the importance score of landmark i is

$$s_i = \frac{1}{NT'} \sum_{n=1}^{N} \sum_{t=1}^{T'} m_{i,t}^{(n)}. \tag{2}$$

Given retention percentage $p \in \{20, 30, 40, 50, 100\}$ and L landmarks, $K = \max(1, \mathrm{round}(Lp/100))$ landmarks with the highest s_i are retained ($\mathcal{S}_p$); $p = 100$ retains all landmarks as a no-pruning baseline. The score is computed using only the inner-training partition of the current fold (Section 3.6), so $\mathcal{S}_p$ is recomputed independently per fold and seed, avoiding any leakage from internal-validation or test samples into landmark selection.

3.3. Input Modalities

Three modalities are constructed from $\mathcal{S}_p$: *xy* (geometric position only, two channels), *optical_flow* (motion only, two channels), and *combined* (concatenation, four channels). The resulting tensor has shape (T', K, C) with $C \in \{2, 2, 4\}$. All three modalities are evaluated under every retention level.

3.4. CNN-GRU Architecture

The (T', K, C) tensor is processed in two stages: a per-frame convolutional extractor (four 1D convolutional blocks over the K-landmark axis, filters $(16, 32, 64, 32)$, kernels $(9, 7, 5, 3)$, each with batch normalization and ReLU, max-pooling after blocks 2 and 3, followed by global average pooling), applied with shared weights across all T' frames; and a recurrent stage over the resulting T' per-frame embeddings. A GRU layer (32 units, recurrent dropout 0.3) returns the full sequence $h_1, \ldots, h_{T'}$ following the standard GRU update [1], which is then pooled by temporal attention:

$$\alpha'_t = \tanh(w_a^\top h_t + b_a), \quad \alpha_t = \frac{\exp(\alpha'_t)}{\sum_\tau \exp(\alpha'_\tau)}, \quad c = \sum_t \alpha_t h_t, \tag{3}$$

producing a sequence-level context vector c. After dropout, c passes through a 16-unit ReLU dense layer with L_2 regularization, batch normalization, and a final softmax layer over the six emotion classes.

3.5. Class-Imbalance Handling via BorderlineSMOTE

BorderlineSMOTE [10] restricts synthetic-sample generation to minority-class examples near the decision boundary, identified by inspecting each example's m nearest neighbors across all classes; only borderline examples, together with k same-class neighbors, generate synthetic examples by interpolation, extending base SMOTE [11]. It is applied only within the inner-training partition, after the internal-validation carve-out and before model fitting (never on internal-validation or test data), on the flattened $(T' \cdot K \cdot C)$-dimensional feature vector, with $k = \min(5, c_{\min} - 1)$ and $m = \min(10, c_{\min} - 1)$ where $c_{\min}$ is the fold's smallest inner-training class count;

random oversampling is used as a fallback when $c_{\min} - 1 < 1$. This directly follows the caution in Section 2.2 that oversampling before, rather than within, a split can leak test information [30].

3.6. Leakage-Free Training and Evaluation Procedure

An outer, subject-level N_{splits}-fold split partitions subjects so each appears in exactly one test fold. Within each fold's training portion, an inner, subject-disjoint split carves out an internal-validation subset governing early stopping and learning-rate scheduling, leaving an inner-training subset for landmark scoring, resampling, and training. The test fold is never used for any decision until final, one-time evaluation. Algorithm 1 states this formally.

Algorithm 1: Leakage-free training and evaluation procedure, one fold

Input: landmark sequences X, labels y, subjects g; retention p, modality mode, resampling flag smote; validation fraction f_{val}, seed.
Output: test-fold predictions and metrics.
Obtain outer training/test subject split $I_{\text{train}}, I_{\text{test}}$ (subject-disjoint);
Split I_{train} by subject into $I_{\text{inner}}, I_{\text{val}}$ using f_{val} and seed (subject-disjoint);
Compute flow features (Eq. 1) for $I_{\text{inner}}, I_{\text{val}}, I_{\text{test}}$;
Compute landmark scores s_i (Eq. 2) using I_{inner} only; select $\mathcal{S}_p$;
Build modality-encoded tensors for all three subsets using $\mathcal{S}_p$;
if smote **then**
  Apply BorderlineSMOTE (or fallback) to I_{inner} only
else
  leave I_{inner} unchanged
Initialize CNN-GRU with weights seeded by seed;
while *stopping criterion not met* **do**
  update weights on I_{inner}; evaluate on I_{val}; update LR/early-stop counter using I_{val} accuracy
Restore best-I_{val}-accuracy weights; predict on I_{test} exactly once; compute metrics;
return *test-fold predictions and metrics*

Weights are updated with Adam under sparse categorical cross-entropy, with gradient-norm clipping. Early stopping and learning-rate reduction both monitor internal-validation, not test, accuracy, which is the specific mechanism preventing test-set-as-validation leakage (Section 2.3). To confirm explicitly: Algorithm 1 applies BorderlineSMOTE (line 6) to I_{inner} only, strictly after the subject-disjoint inner/validation split (line 2); I_{val} is never passed to the resampling step and is used, unresampled, solely to compute the early-stopping and learning-rate-reduction criterion. The same holds for I_{test}, which is untouched by resampling and by every training-time decision, consistent with the one-time evaluation stated above.

4. Experimental Setup

4.1. Datasets

CK+ [37] contributes 309 sequences from 106 subjects across the six basic classes (excluding contempt): 45 anger, 59 disgust, 25 fear, 69 happiness, 28 sadness, 83 surprise, matching the CK+ paper's own accounting after excluding contempt. IMED [15] contributes 830 samples from 15 subjects, restricted to the same six classes (excluding its twelve mixed-emotion categories and neutral class): 155 anger, 133 disgust, 146 fear, 127 happiness, 172 sadness, 97 surprise. IMED's label order was confirmed by the collaborator who performed its extraction to match CK+'s order; this is verbal, not code-verified as for CK+. Both datasets use identical preprocessing: MediaPipe Face Mesh (v0.10.35) landmarks, x, y position only (CK+'s additional depth coordinate is discarded), and a deterministic temporal subsampling to $T = 10$ frames – if the raw sequence has fewer than T frames the

last frame is repeated, otherwise frame $h \in \{1, \ldots, T\}$ is drawn from position $N_i - (T-h)K - 1$ with stride $K = \lfloor N_i/T \rfloor$, concentrating sampled frames toward the sequence's later, apex-ward portion. Table 1 summarizes both datasets; IMED's much smaller subject pool relative to its sample count is the specific structural contrast RO3 probes.

Table 1. Dataset summary.

	CK+	IMED
Samples / Subjects / Classes	309 / 106 / 6	830 / 15 / 6
Frames / Landmarks per frame	10 / 468	10 / 468
Anger/Disgust/Fear	45/59/25	155/133/146
Happiness/Sadness/Surprise	69/28/83	127/172/97

4.2. Data Splitting, Experimental Grid, and Hyperparameters

Both datasets use $N_{\text{splits}} = 10$ outer, subject-level GroupKFold folds and an internal-validation fraction $f_{\text{val}} = 0.15$ via subject-level GroupShuffleSplit; the outer split is deterministic and identical across seeds. Every combination of five retention levels, three modalities, two resampling conditions, five seeds $(42, 84, 126, 168, 210)$, and ten folds is evaluated per dataset: $5 \times 3 \times 2 \times 5 \times 10 = 1500$ fold-level results each. Prior to testing, each merged result set was checked for duplicate or missing configuration-fold-seed combinations, nulls, and internal subject-count consistency; both datasets passed all four checks. Table 2 lists the fixed hyperparameters, shared across both datasets and the full grid, with no per-dataset or per-configuration tuning.

Table 2. Training hyperparameters.

Hyperparameter	Value
Optimizer / Initial LR / Grad. clip	Adam / 1×10^{-3} / 1.0
Batch size / Max epochs	16 / 150
Early stop (monitor, patience)	internal-val. accuracy, 30 epochs
LR reduction (factor, patience, min)	0.5, 10 epochs, 1×10^{-5}
L_2 / Dropout / GRU rec. dropout	3×10^{-4} / 0.5 / 0.3

4.3. Software, Hardware, and Metrics

Experiments ran on Kaggle Notebooks (Python 3.12.13, single NVIDIA T4 GPU, CUDA 13.0) using TensorFlow 2.20.0, scikit-learn 1.6.1, imbalanced-learn 0.14.2, pingouin 0.6.1, scikit-posthocs 0.14.0, and MediaPipe 0.10.35. Weighted F1 (`test_f1_weighted`) is the primary metric, since accuracy can mask minority-class performance under imbalance and weighted F1 better reflects overall performance under the natural class distribution than macro F1; accuracy is reported as a secondary robustness check.

4.4. Statistical Testing Procedure

For each dataset, testing uses fold as the unit of replication, on fold-level means (averaged across the five seeds first, so correlated seed-fold points do not inflate the effective sample size). The main test applies a Friedman test across all 30 configurations ($n = 10$ folds), followed by Nemenyi post-hoc comparisons; 95% confidence intervals for each configuration use both a paired t-interval ($\mathrm{df} = 9$) and a 10,000-resample percentile bootstrap. Three marginal tests then isolate each factor by pooling over the other two: a paired Wilcoxon signed-rank test for resampling, and a Friedman test with Nemenyi post-hoc for modality and for landmark retention, each with 95% t-interval confidence intervals for headline pairwise differences; all three are repeated with accuracy as a robustness check. Kendall's W is interpreted on the scale of Landis and Koch [38]: slight (W<0.20), fair (0.20–0.40), moderate (0.40–0.60), substantial (0.60–0.80), almost perfect (≥ 0.80).

Beyond per-test $\alpha = 0.05$, two further analyses are reported. A Bonferroni and a Benjamini-Hochberg FDR correction are applied within each dataset's family of four headline tests (main + three marginal), with both corrected and uncorrected decisions reported. Because CK+'s and IMED's outer folds are drawn from disjoint, unmatched subject pools, a Dataset-by-Factor interaction cannot be tested within a single blocked model; instead, a fold-level effect size is computed within each dataset for each factor (combined−xy for modality, 30%−100% for retention, SMOTE−no-SMOTE for resampling, each pooled over the other two factors and all seeds), and the two datasets' ten effect sizes are compared with an unpaired, two-sided Mann-Whitney U test; the resulting three p-values are themselves corrected (Bonferroni and FDR).

5. Results

Results are reported for CK+ (5.1) and IMED (5.2) separately rather than pooled, since RO3 asks whether effects on one dataset generalize to the other. All comparisons use weighted F1 as the primary metric. Table 3 collects every headline test statistic in one place before the detailed narrative; the full 30-by-30 Nemenyi matrix underlying the main-omnibus row of each dataset is reproduced in part in Appendix 7, restricted to comparisons against each dataset's best configuration.

Table 3. Summary of headline test statistics, both datasets.

Dataset	Test	Statistic	Effect size
CK+	Main omnibus	Friedman, $Q = 215.997$ (df 29)	$W = 0.745$ (subst.)
CK+	Mode	Friedman, $Q = 15.8$ (df 2)	$W = 0.79$ (subst.)
CK+	TopK	Friedman, $Q = 37.6$ (df 4)	$W = 0.94$ (a. perfect)
CK+	SMOTE	Wilcoxon, $W_{\text{stat}} = 0$	RBC $= 1.0$
IMED	Main omnibus	Friedman, $Q = 51.649$ (df 29)	$W = 0.178$ (slight)
IMED	Mode	Friedman, $Q = 4.2$ (df 2)	$W = 0.21$ (fair)
IMED	TopK	Friedman, $Q = 13.44$ (df 4)	$W = 0.336$ (fair)
IMED	SMOTE	Wilcoxon, $W_{\text{stat}} = 7.0$	RBC $= 0.745$

5.1. CK+ Results

The omnibus Friedman test across all 30 configurations is significant ($p = 1.7 \times 10^{-30}$), indicating highly consistent ranking across folds (Figure 2, left). The best configuration is top30_combined_smoteTrue, mean weighted F1 $= 0.810$ (95% CI $[0.779, 0.841]$), mean rank 2.3/30. Nemenyi finds 11 of the remaining 29 configurations significantly worse ($p \leq 0.05$), concentrated among 100%-retention, single-modality, no-resampling combinations; the remaining 19 are not distinguishable from the best at $n = 10$.

Landmark retention is significant ($p = 1.4 \times 10^{-7}$); mean F1w decreases monotonically from 20% (0.740) through 30% (0.725), 40% (0.700), 50% (0.686) to 100% (0.616) (Figure 3, left). The 30-vs-100% CI $[0.091, 0.128]$ and 20-vs-50% CI $[0.039, 0.069]$ exclude zero; the 20-vs-30% CI $[-0.0001, 0.029]$ includes zero by a narrow margin, so 20% and 30% are not reliably distinguishable from each other, though both are reliably better than 50% or 100%.

Modality is significant ($p = 3.7 \times 10^{-4}$): combined (0.748) outperforms xy (0.657) and optical_flow (0.675), CIs $[0.074, 0.107]$ and $[0.044, 0.103]$ excluding zero (Figure 4, left); xy-vs-optical_flow CI $[-0.046, 0.012]$ includes zero.

SMOTE is significant ($p = 0.00195$, RBC$= 1.0$, 10/10 folds); mean F1w 0.724 with SMOTE vs. 0.662 without, diff. 0.062, CI $[0.042, 0.082]$ (Figure 5, left); the accuracy robustness check agrees qualitatively.

Landmark selection frequency (Figure 6, top, CK+ at 30%) is strongly bimodal: most indices are selected in nearly all or almost no seed-fold combinations, clustering in contiguous index ranges rather than scattering uniformly, indicating a stable selected subset across training subsets. Appendix 7 maps named MediaPipe facial regions to their index sets for readers cross-referencing this figure against facial anatomy.

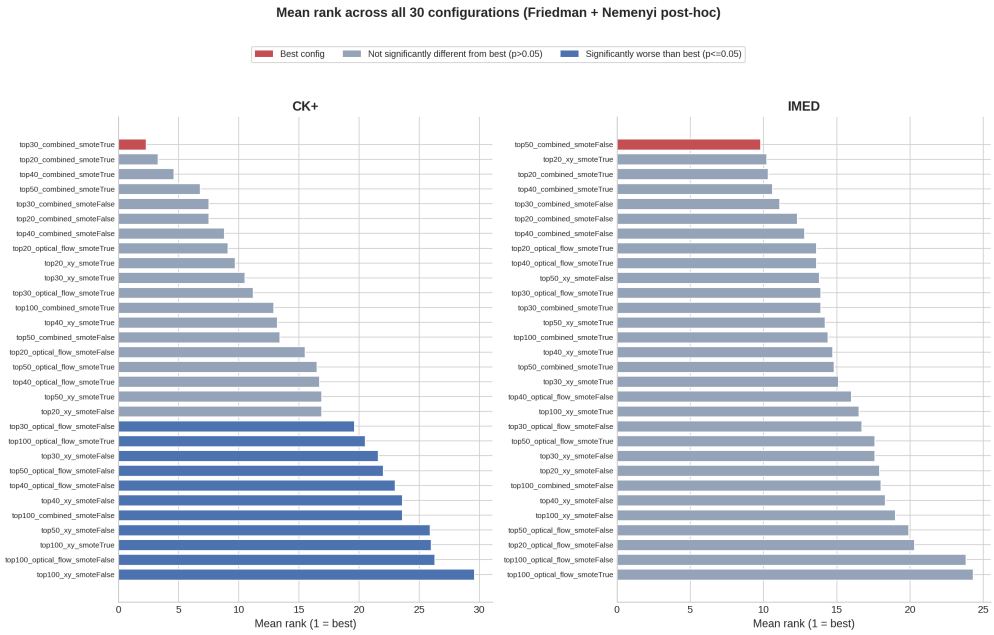


Figure 2. Mean rank across all 30 configurations, both datasets.

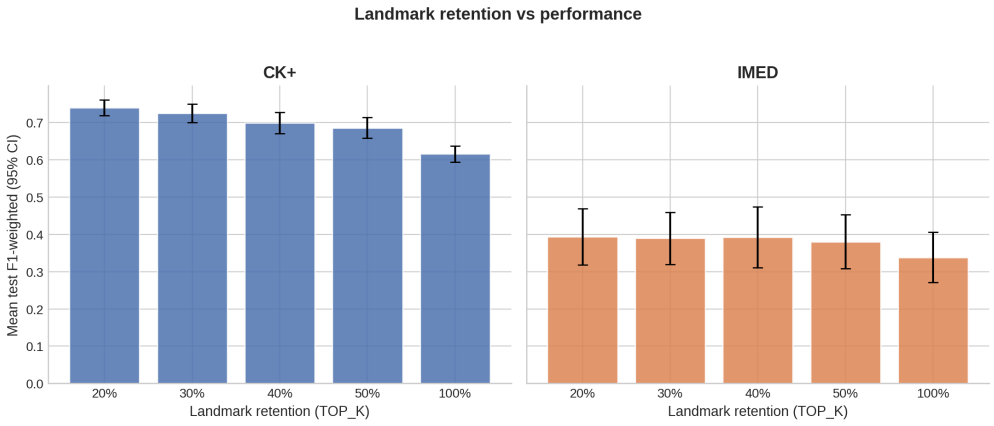


Figure 3. Landmark retention (Top-K) versus performance, both datasets.

Figure 7 answers the anatomical-mapping question directly by plotting selection frequency at each index's canonical position rather than leaving readers to cross-reference index numbers against Appendix 7 by hand. Selection concentrates overwhelmingly on the **lips** and **eyebrows** (near-100% selection frequency across folds and seeds at both datasets' headline retention levels), is negligible on the eye region itself, and shows light, inconsistent coverage of the forehead/hairline and jawline. For CK+ (top30%), 147 of 468 indices are ever selected across the 50 runs, with a stable core of approximately 140 (matching the fixed top30%≈140/468 retention size), indicating a highly consistent selected subset fold-to-fold; for IMED (top50%), 254 of 468 indices are ever selected, with more fold-to-fold variability, consistent with IMED's smaller per-fold subject pool. This lips-and-eyebrows concentration is consistent with FER literature identifying mouth and brow movement as the dominant discriminative cues for the six basic emotions tested here, and supports the validity of the flow-derived importance score: the algorithm selects landmarks previously established as informative rather than scattering across the face.

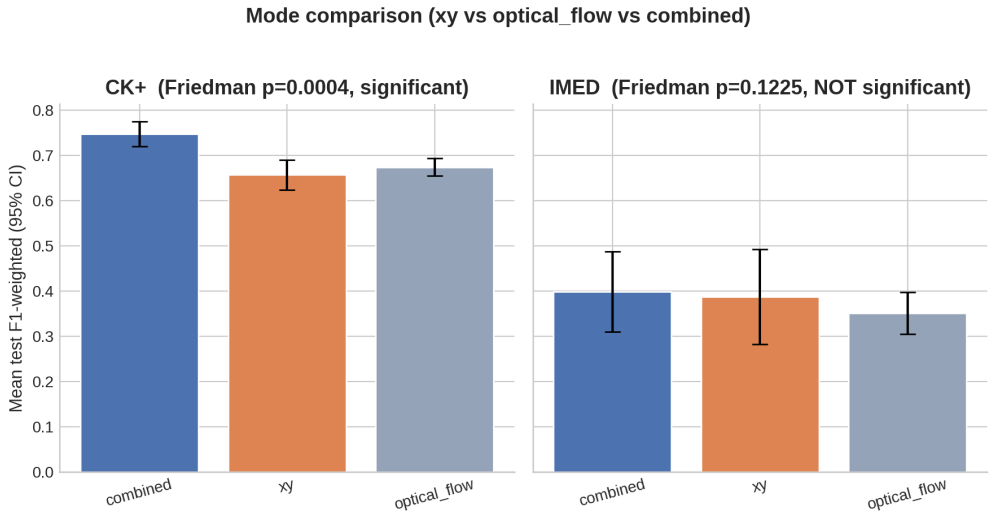


Figure 4. Input modality comparison, both datasets.

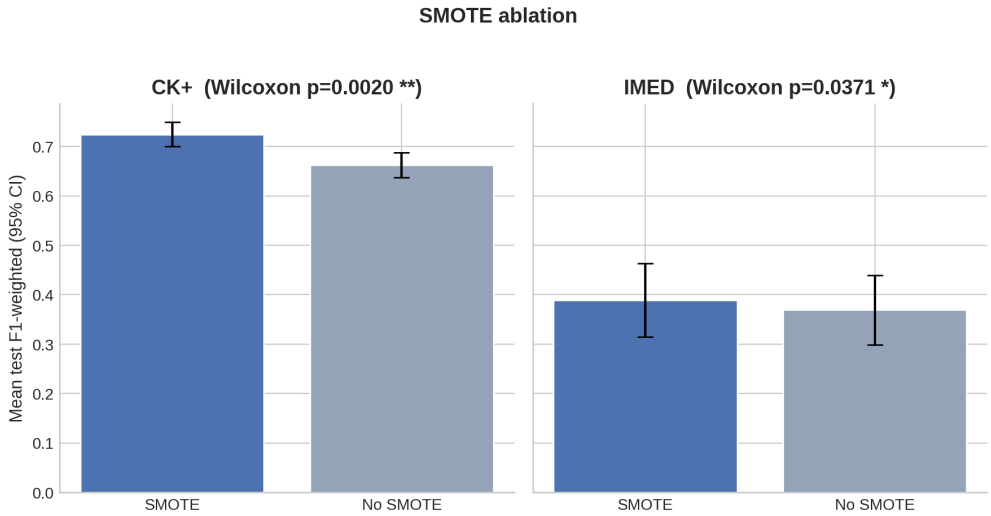


Figure 5. BorderlineSMOTE ablation, both datasets.

For CK+ (Figure 8, left), per-class recall is highest for surprise (0.98) and happiness (0.96), lowest for anger (0.56), most confused with disgust (0.28) and sadness (0.13); sadness is most confused with anger (0.25).

For CK+ (Figure 9, left), training accuracy rises steadily above 0.95 while internal-validation plateaus, with fold-to-fold noise, around 0.85–0.90; losses converge by roughly epoch 80.

Table 4 (pooled over folds/seeds, matching Figure 8's aggregation, which generally differs slightly from the fold-level-mean metric used for hypothesis testing) shows CK+'s per-class F1 ranges 0.573–0.981; surprise is best-recognized on both datasets, anger and its confusable neighbors are weakest.

5.2. IMED Results

The omnibus test is significant but with a substantially smaller effect ($p = 0.0060$, $W = 0.178$ vs. 0.745 for CK+). The best configuration is top50_combined_smoteFalse, mean F1w $= 0.426$ (CI $[0.311, 0.542]$), mean rank 9.8/30 –

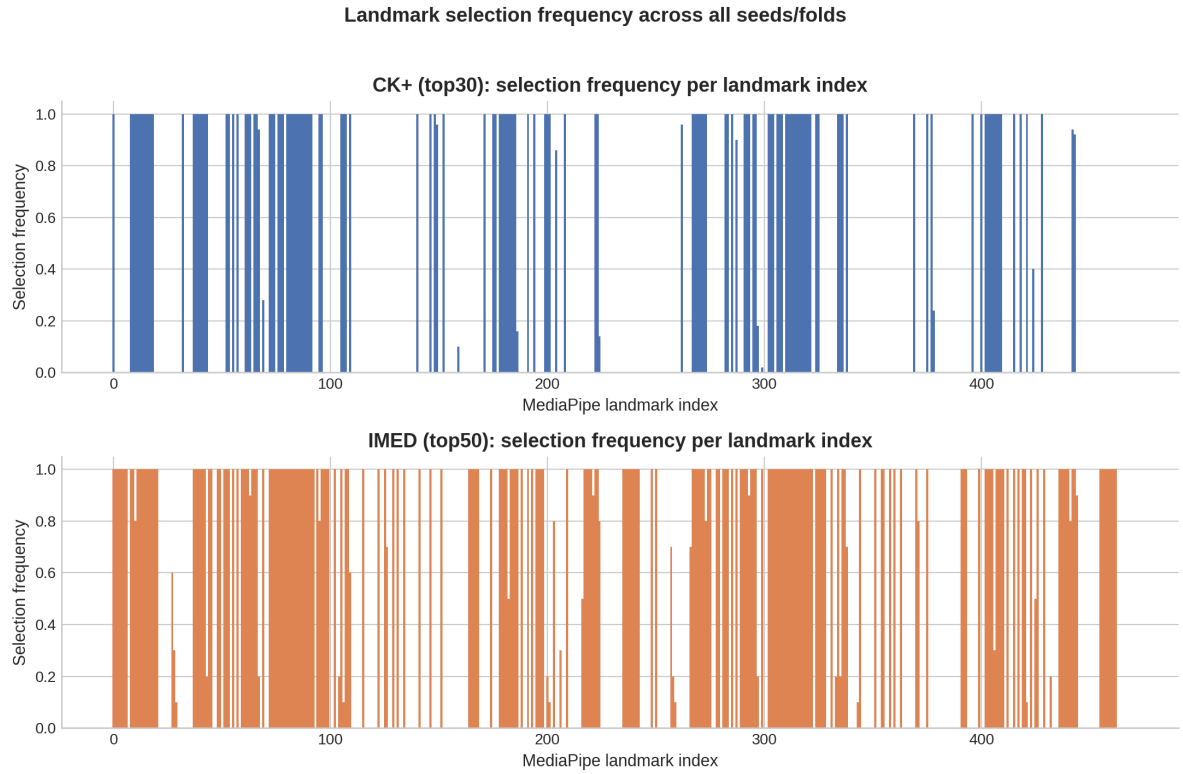


Figure 6. Landmark selection frequency across seeds and folds, both datasets.

Table 4. Per-class precision, recall, and F1 for each dataset's best configuration, pooled over folds and seeds.

	CK+ (top30, combined, SMOTE)			IMED (top50, combined, no SMOTE)		
Class	Prec.	Rec.	F1	Prec.	Rec.	F1
Anger	0.592	0.556	0.573	0.505	0.432	0.466
Disgust	0.732	0.770	0.750	0.326	0.311	0.319
Fear	0.840	0.672	0.747	0.399	0.288	0.334
Happiness	0.880	0.957	0.917	0.493	0.699	0.579
Sadness	0.681	0.671	0.676	0.523	0.535	0.529
Surprise	0.985	0.976	0.981	0.614	0.689	0.649
Macro avg	0.785	0.767	0.774	0.477	0.492	0.479

notably not the top30/SMOTE-enabled pattern best on CK+. Nemenyi finds zero of the remaining 29 configurations significantly different from the best (Figure 2, right); every configuration is statistically indistinguishable from the best at this sample size.

Landmark retention remains significant ($p = 0.0093$, $W = 0.336$) but far flatter: 20% (0.394), 30% (0.390), 40% (0.392), 50% (0.381), dropping only at 100% (0.339) (Figure 3, right). Only the 100%-vs-others comparisons exclude zero; 20/30/40/50% are mutually indistinguishable. Modality is not significant ($p = 0.122$, $W = 0.21$): combined (0.399), xy (0.387), optical_flow (0.351), same order as CK+ but every pairwise CI includes zero (Figure 4, right) – contrasting directly with CK+'s significant modality effect. SMOTE remains significant but more weakly ($p = 0.0371$, RBC$= 0.745$); mean F1w 0.389 with SMOTE vs. 0.369 without, diff. 0.020, CI $[0.002, 0.038]$, roughly a third the CK+ interval's width (Figure 5, right); the accuracy check is borderline ($p = 0.049$).

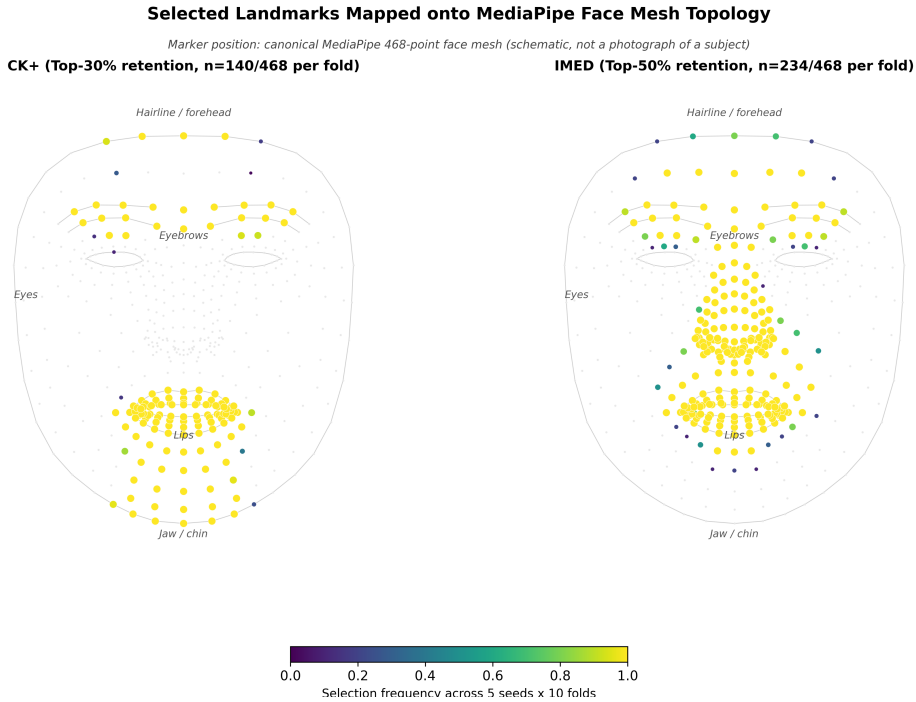


Figure 7. Selected landmarks mapped onto the MediaPipe canonical face-mesh template. Selection frequency per index is the fraction of the 50 fold-seed runs (headline configuration) in which that index appears in the selected set, plotted at that index's canonical (x, y) position (geometric template only, no photograph or identity).

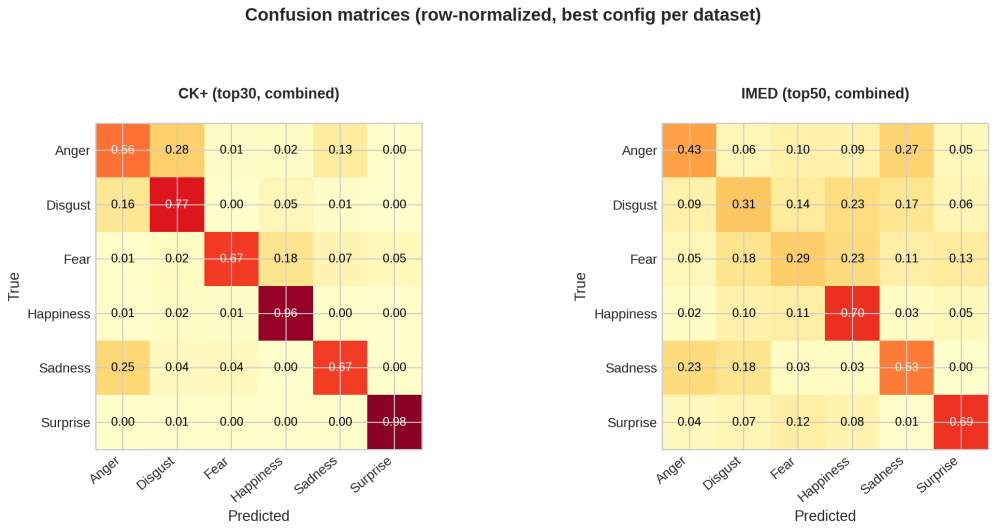


Figure 8. Row-normalized confusion matrix for the best configuration, both datasets.

IMED's single best configuration uses no SMOTE, which is not a contradiction of the marginal SMOTE finding: the marginal test pools 750 fold-level comparisons into ten paired means, far more power than any single configuration comparison, and the omnibus test found zero of 29 configurations distinguishable from the best, so



Figure 9. Representative training and internal-validation curves, best configuration per dataset.

the identity of the nominal best is within the range expected from sampling noise at this level of ambiguity – the pooled marginal test, not the single best cell, is the trustworthy source for the resampling conclusion here.

Landmark selection frequency at 50% retention (Figure 6, bottom) shows the same bimodal, contiguously clustered pattern as CK+, indicating selection stability is not specific to the larger subject pool. The confusion matrix (Figure 8, right) shows lower, more evenly spread recall (0.29–0.70); disgust is confused with fear, happiness, and sadness simultaneously rather than concentrating on one class; the anger-sadness bidirectional confusion pattern recurs (anger→sadness 0.27, sadness→anger 0.23). Training dynamics (Figure 9, right) show comparable training accuracy (0.85–0.90) but internal-validation plateauing much lower (0.45–0.55) with visibly larger fluctuation and loss spikes above 3.0, not present for CK+. Table 4 shows IMED's per-class F1 ranging 0.319–0.649, a smaller absolute span than CK+ but a much lower ceiling.

IMED's fifteen subjects each contribute enough pooled test instances (95–545, over the 5 seeds $\times$ 10 folds) for a well-powered per-subject accuracy estimate, unlike CK+, where 106 subjects average only 1–6 pooled instances each and a per-subject breakdown would mostly reflect single-sample noise rather than a stable subject effect; the per-subject view is therefore reported for IMED only. Table 5 shows substantial subject-to-subject variation, from 20.7% to 75.3% accuracy.

Five subjects (001, 002, 011, 012, 013) fall below 36% accuracy and jointly account for a disproportionate share of all pooled test instances; Subject001 alone contributes 545 of 4,150 pooled predictions ($\approx$13%). This is a well-powered, subject-level finding, not a small-sample artifact, and it identifies a specific, quantified driver of IMED's lower aggregate ceiling: a handful of subjects pull the dataset-level mean down substantially more than a uniformly-distributed error rate would predict. Section 6.5 returns to the practical implication of this concentration.

Table 5. IMED per-subject accuracy, headline configuration (top50, combined, no SMOTE), pooled over 5 seeds × 10 folds.

Subject	n (pooled)	Accuracy
Subject001	545	0.207
Subject012	170	0.253
Subject013	95	0.295
Subject011	190	0.353
Subject002	445	0.353
Subject005	315	0.467
Subject015	160	0.519
Subject014	170	0.535
Subject003	515	0.540
Subject006	345	0.548
Subject010	175	0.600
Subject008	290	0.634
Subject009	290	0.645
Subject007	275	0.691
Subject004	170	0.753

5.3. Comparison with Prior Work

Table 6 places this study's results alongside eight prior CK+-evaluating studies and two prior IMED-evaluating studies, each verified directly against its source publication.

Table 6. Comparison with prior work on CK+ and IMED.

Method	**Dataset (subjects)**	**Classes**	**Subj.-disjoint**	**Acc.**
Res-RBG [2]	CK+ (118)	7	Not stated	98.10%
Cascade-attn. [3]	CK+ (not stated)	7	Not stated	99.23%
SACNN-ALSTMs [4]	CK+ (118)	7	Not stated	95.15%
Delaunay+QDA [17]	CK+ (123)	7	Not stated	100%
Landmark ROI-CNN [18]	CK+ (123)	6	Yes	98.04%
Improved VGG19 [33]	CK+ (123)	7	Not stated	93.94%
ResNet50-GRU [34]	CK+ (not stated)	7	Not stated	95.56%
ERFN [23]	CK+ (118)	6	Not stated	98.00%
This study (top30, comb., SMOTE)	CK+ (106)	6	Yes	81.0%
Stacked CNN [35]	IMED+CK+FER2013 (15/123/–)	6/7	Not stated	71.0%
ViT/Swin/ConvNeXt [36]	IMED (15)	18 (12 mixed + 6 basic)	No (image-level)	Not comparable*
This study (top50, comb., no SMOTE)	IMED (15)	6	Yes	42.6%

*Akeh and Kusuma [36] evaluate an eighteen-category label set (IMED's twelve mixed-emotion categories plus six basic-emotion categories, classified jointly) rather than the six basic-emotion classes used here and in Mahastama et.al [35], reporting 79.37% test accuracy with their best architecture (Vision Transformer). This figure is omitted from the Acc. column rather than placed alongside the six-class figures above, since the two are not measuring the same classification task, and since their randomly shuffled, class-stratified image-level split, with no stated subject-level grouping across IMED's fifteen subjects, makes the reported figure additionally vulnerable to the identity-leakage inflation this study's own protocol is designed to avoid (Section 2.3).

Every prior CK+ result exceeds this study's F1w of 81.0%, range 93.94–100%. Three structural differences separate this from a controlled benchmark. First, seven of nine prior methods do not state a subject-disjoint protocol; Ayeche and Alti report exactly 100% under an 80/20 split with no subject grouping [17], while Soylemez and Ergen, the only subject-disjoint study here, reports 98.04%, in the same range rather than visibly lower – consistent with, though not proof of, inflation in the others [18]. Second, six of nine use a seven-class labeling including contempt from a differently filtered pool (118 or 123 subjects; two studies state no count), versus this

study's six-class, 106-subject subset; a protocol-matched re-implementation would need to standardize the class set too. Third, two prior IMED studies exist: Mahastama evaluated on three combined datasets rather than IMED alone under an unstated subject-grouping protocol [35], and Akeh evaluated jointly on IMED's twelve mixed-emotion and six basic-emotion categories (eighteen classes total) under a class-stratified but not subject-disjoint image-level split, rather than the six-class basic-emotion subset used here [36]; neither offers a directly comparable, subject-disjoint, six-class IMED baseline.

5.4. Multiple-Comparison Correction and Cross-Dataset Interaction Tests

Table 7. Multiple-comparison correction across the four headline tests per dataset.

Dataset	Test	Raw p	Bonferroni	FDR	Sig. (FDR)
CK+	Main omnibus	1.7×10^{-30}	7.0×10^{-30}	7.0×10^{-30}	Yes
CK+	Mode	0.000371	0.001483	0.000494	Yes
CK+	TopK	1.4×10^{-7}	5.4×10^{-7}	2.7×10^{-7}	Yes
CK+	SMOTE	0.001953	0.007813	0.001953	Yes
IMED	Main omnibus	0.005964	0.023855	0.018629	Yes
IMED	Mode	0.122456	0.489826	0.122456	No
IMED	TopK	0.009314	0.037258	0.018629	Yes
IMED	SMOTE	0.037109	0.148438	0.049479	Yes

Every CK+ conclusion is unchanged by either correction. On IMED, only the resampling test is correction-sensitive: it survives FDR (adj. $p = 0.049$) but not Bonferroni (adj. $p = 0.148$).

Table 8. Dataset-by-Factor interaction tests (Mann-Whitney U, fold-level effect sizes).

Factor (effect)	CK+ mean	IMED mean	Raw p	Bonferroni
Mode (combined − xy)	0.091	0.012	0.0113	0.0340
TopK (30% − 100%)	0.109	0.051	0.0017	0.0051
SMOTE (True − False)	0.062	0.020	0.0058	0.0174

All three factors have significantly larger effect size on CK+ than IMED, robust to Bonferroni correction across the three-test family; every point estimate is smaller on IMED, ranging from roughly one-half (retention, resampling) to one-eighth (modality) of the CK+ magnitude, foreshadowing the interpretation in Section 6.3.

5.5. Computational Cost

Table 9 reports training and inference cost for each dataset's headline configuration (CK+: top30%, combined, SMOTE; IMED: top50%, combined, no SMOTE), aggregated over the same 50 fold-seed runs (5 seeds $\times$ 10 folds) as every other result in this section, on the Kaggle Notebooks / single NVIDIA T4 GPU environment stated in Section 4.3.

Table 9. Computational cost per fold-seed run, headline configurations ($n = 50$ runs each).

Dataset	Epochs (mean ± SD)	Train time, s (mean ± SD)	Infer. time, s (mean ± SD)	Infer., ms/sample
CK+	91.0 ± 14.6	127.8 ± 21.2	1.25 ± 0.07	40.5 ± 2.1
IMED	84.7 ± 22.2	521.6 ± 135.3	2.46 ± 0.07	30.5 ± 5.3

IMED's longer per-run training time reflects its larger sample count (830 versus 309) and larger retained landmark set at these headline retention levels (top50% versus top30%), not an architectural difference: both datasets use the identical CNN-GRU architecture and hyperparameter grid, with no per-dataset tuning (Section 4.2).

The longest single run observed was 205.6s (CK+) and 1065.2s (≈17.8 minutes, IMED), both comfortably within a single Kaggle session. Per-sample inference cost is low for both datasets (31–41 ms/sample), indicating the trained model is lightweight enough for near-real-time application despite the accuracy limitations discussed in Section 6.3.

6. Discussion

This section interprets Section 5 with respect to RO1 (6.1), RO2 (6.2), and RO3 (6.3), states the study's strengths (6.4), and its limitations (6.5). Every claim below is restricted to the two datasets, six classes, and grid actually tested.

6.1. RO1: Landmark Selection and Modality Fusion

Both datasets agree that retaining all 468 landmarks is reliably worst (CIs exclude zero for 30-vs-100% and 20-vs-100% on both), consistent with the premise that many landmarks, e.g. on the outer contour, carry little discriminative motion, and directionally consistent with Soylemez s 84-to-98.04% gain from landmark-ROI restriction under a subject-disjoint protocol [18]'. They disagree on the finer 20-vs-30% ordering, which is unresolved on either dataset at this resolution.

On modality, the two datasets diverge qualitatively rather than by degree: combined fusion is significantly better on CK+ but not distinguishable from single modalities on IMED, despite the same numerical ordering. Section 6.3 unifies this with the corresponding landmark-retention pattern via the formal interaction test, rather than treating modality's failure to replicate as an isolated finding. A caveat applies to literature comparisons in Table 6: none of the cited studies confirm a subject-disjoint protocol except Soylemez [18], so their figures and this study's weighted F1 under a leakage-aware protocol are not directly comparable quantities.

Beyond protocol differences, a further question is whether the CK+ gap against prior work partly reflects model capacity rather than the leakage-free protocol alone: the CNN-GRU architecture used here is comparatively shallow next to more recent macro-FER architectures, such as Vision Transformers or deeper 3D-CNNs, none of which were evaluated in this study. This study cannot separate a protocol effect from an architecture-capacity effect using the experiments already run, since both factors differ simultaneously between this work and the cited state-of-the-art figures, and no architecture ablation under the identical subject-disjoint protocol was performed. Whether a higher-capacity or attention-based architecture narrows the gap under this study's own leakage-free protocol is left to future work (Section 7) rather than claimed or ruled out here.

6.2. RO2: Class-Imbalance Handling via BorderlineSMOTE

The marginal Wilcoxon test supports SMOTE on both datasets (CK+ 10/10 folds, $p = 0.00195$; IMED majority of folds, $p = 0.0371$), with both CIs excluding zero, though IMED's interval is a third the width of CK+'s and its accuracy-based check is borderline ($p = 0.049$). As Table 7 shows, IMED's SMOTE result is the one finding in this study whose significance is correction-method-dependent (FDR yes, Bonferroni no), reported rather than resolved by selecting the favorable correction. This extends Borderline-SMOTE, previously validated in FER only as one factor among several in bundled pipelines [39], as an isolated, controlled resampling ablation.

IMED's single best configuration uses no SMOTE, which is not a contradiction: the marginal test's far greater statistical power (750 pooled comparisons vs. a single configuration among 30 that Nemenyi could not distinguish from the best at all) makes the pooled test the trustworthy source, not the identity of the nominally best cell.

6.3. RO3: Cross-Dataset Generalization and Dataset-Scale Sensitivity

Table 8's interaction tests, rather than a side-by-side reading of separate tests, give the direct answer: all three factors' effect sizes are significantly smaller on IMED than CK+, robust to correction. No effect reverses direction; every one shrinks, modality most severely (roughly one-eighth of its CK+ magnitude) against roughly one-half for retention and resampling. This reframes Section 5's pattern precisely: retention and resampling remain significant

within IMED because their CK+ margins were large enough to absorb a ∼50% reduction; modality does not, not because it differs qualitatively, but because its attenuation is more severe and its starting margin narrower.

CK+'s 106 subjects average ∼10–11 per test fold versus IMED's 15 subjects averaging only 1–2, so an IMED fold-level estimate is dominated by one or two individuals' idiosyncratic style, inflating between-fold variance for every factor simultaneously – reflected in every IMED Kendall's W dropping at least one full agreement category below its CK+ counterpart (Table 3). This is consistent with evidence that small datasets amplify estimate variance and instability [31], and with Li and Deng's demonstration that high within-dataset FER accuracy does not guarantee generalization, reinforcing why a controlled replication attempt, rather than treating CK+ alone as conclusive, is necessary [14]. Because both datasets were evaluated under an identical protocol, this pattern is attributable to the datasets themselves rather than to differing evaluation rigor – the specific guarantee RO3 was designed to provide. Practically: landmark pruning and BorderlineSMOTE resampling should be expected to weaken but often remain useful on a much smaller target dataset, while modality-fusion benefit should not be assumed to transfer without direct validation.

6.4. Strengths

Four contributions extend beyond the accuracy figures themselves. First, the identical subject-disjoint internal-validation protocol (Algorithm 1) is applied to both datasets rather than relaxed for either; Section 2.3 shows several cited CK+ studies, including close architectural analogs, do not state whether their own protocol is subject-disjoint. Second, the statistical design (seed-averaging before testing, fold as replication unit, per-dataset rather than pooled reporting) avoids pseudo-replication and premature generalization; the RO3 finding is only visible because of this design. Third, BorderlineSMOTE is evaluated as an isolated ablation rather than a bundled pipeline component, directly addressing the RO2 gap. Fourth, Table 6 is published in full even though every prior CK+ figure exceeds this study's own, with the gap explained by verifiable protocol and class-set differences rather than omitted or reframed – a transparency that lets a reader judge how much of the gap reflects the leakage-aware protocol versus a genuine capability difference.

6.5. Limitations

Several scope decisions share one resource constraint: all experiments ran on Kaggle Notebooks under a 12-hour session limit, and the full grid already represents a substantial share of the compute available to a five-person team across five accounts. This underlies four related limitations: class-imbalance handling is evaluated as BorderlineSMOTE versus no resampling only, not the broader landscape (focal loss, ADASYN, etc.); the internal-validation fraction is a single 15% GroupShuffleSplit rather than nested cross-validation; no hyperparameter tuning was performed; and the same fixed architecture is used for both datasets, a deliberate choice keeping the RO3 comparison attributable to the datasets rather than per-dataset tuning, but meaning neither dataset's performance reflects a dataset-specific search.

The same resource constraint left several further comparisons unaddressed in this revision, named explicitly rather than left implicit. The frequency with which BorderlineSMOTE's random-oversampling fallback is triggered, used when a minority class has too few neighbors for borderline interpolation, was not logged per fold during the original experiment runs and cannot be recovered without re-instrumenting and re-running the pipeline. An ablation over alternative frame-sampling strategies (uniform, motion-peak, fixed early-middle-late sampling), alternative modality-fusion mechanisms (attention or gated fusion) beyond the concatenation used here, an alternative landmark-importance score (e.g., class-separability- or classifier-weight-based rather than motion-magnitude-based), and cross-dataset training or fine-tuning between CK+ and IMED were all identified as valuable directions during review but each requires new training runs outside this revision's scope; all are restated as concrete future-work directions in Section 7.

IMED's 15-subject pool (versus CK+'s 106) directly limits what can be concluded from IMED alone: individual folds are small, between-fold variance is high, and null results, notably the non-significant modality effect, should be read as inconclusive at this sample size rather than as evidence the effect is absent. Comparisons to prior work (Table 6) use each source's own reported figures rather than a re-implementation under this study's protocol; all but one of nine prior CK+ studies, and both prior IMED studies, do not state a subject-disjoint protocol, and several use

a different class count from a differently sized subject pool, so a protocol-matched re-implementation was outside the available compute budget.

This has a direct practitioner implication: the per-subject breakdown in Table 5 (Section 5.2) shows five of fifteen IMED subjects below 36% accuracy, jointly dominating the pooled error count, so a practitioner deploying this pipeline on a new, small subject pool should not expect the dataset-level 42.6% weighted F1 to generalize uniformly across individuals; a specific new subject's performance is not predictable from the aggregate figure alone. Because modality fusion did not reach significance on IMED, practitioners working with a similarly small, subject-scarce dataset should not assume the combined-modality configuration is worth its added complexity without validating it directly on their own data, whereas landmark pruning and BorderlineSMOTE resampling, whose IMED effects remained significant despite attenuation, are more defensible defaults to carry over.

Quantifying the underlying power concern: using the marginal SMOTE effect already reported in Section 5.2 (paired Wilcoxon, $n = 10$ folds, mean paired difference 0.0198, 95% paired-t CI $[0.0015, 0.0381]$), a back-calculated Cohen's $d_z \approx 0.775$ implies achieved post-hoc power of only ≈ 0.59 at nominal $\alpha = .05$ for this paired, $n = 10$ design; under this study's own Bonferroni correction ($\alpha = .0125$ for the four-test family, Table 7), achieved power at the observed effect drops to ≈ 0.32, quantitatively consistent with the SMOTE conclusion's own sensitivity to correction method (Section 6.2). The minimum detectable effect for 80% power at $n = 10$ folds is $d_z \approx 1.00$; reaching 80% power for a medium effect ($d_z = 0.5$) would require roughly 34 folds, and for a small effect ($d_z = 0.2$) roughly 199 folds. This d_z is approximated from the paired-t confidence interval via the normal/t-theory relationship to Cohen's d_z, offered because no closed-form post-hoc power formula exists for the Wilcoxon signed-rank test used for the primary inference, and should be read as an approximation rather than an exact nonparametric power calculation. The practical conclusion is that IMED's 15-subject, 10-fold design is adequately powered only for large paired effects, so smaller true effects, including a true modality or retention effect on IMED, cannot be ruled out as false negatives at this sample size; this is an intrinsic limitation of the dataset's public subject count rather than of the analysis pipeline, and resolving it would require more IMED subjects than currently exist in the public release.

Three further limitations concern data provenance and terminology: IMED's label ordering is verbally, not code-verified, confirmed to match CK+'s (the recurring anger-sadness confusion pattern in both datasets is corroborating but not conclusive evidence); the feature called "optical flow" throughout is a finite-difference landmark displacement (Eq. 1), not a pixel-intensity-based flow field, and comparisons to image-level-flow prior work should account for this; and every sequence is subsampled to a fixed ten frames, with variable-length modeling not evaluated. A related limitation is that neither dataset's landmark quality was audited for detector error, occlusion, or head-pose effects: CK+ and IMED are both posed, front-facing, laboratory-recorded corpora, so MediaPipe Face Mesh detection failure is expected to be rare, but this was not measured directly, and the pipeline's robustness to the landmark noise, partial occlusion, and off-axis pose typical of in-the-wild footage is correspondingly untested; extension to in-the-wild data (Section 7) would need to characterize this explicitly rather than assume it carries over. Finally, two statistical gaps flagged during review were addressed rather than left open: multiple-comparison correction (Table 7), which changes only the IMED SMOTE conclusion depending on method, and the formal interaction test (Table 8); the interaction test establishes that effect sizes shrink on IMED but is silent on why beyond the subject-pool-size account above, which remains a plausible, evidence-consistent explanation rather than an isolated causal test.

7. Conclusion and Future Work

This study proposed a CNN-GRU pipeline combining optical-flow-guided Top-K landmark selection, geometric-motion modality fusion, and BorderlineSMOTE, evaluated under an identical, leakage-free, subject-disjoint protocol on two datasets differing by an order of magnitude in subject count. Landmark retention of 20–30% and BorderlineSMOTE each significantly improved weighted F1 on CK+ and, more weakly, on IMED, while modality fusion was significant only on CK+. A direct interaction test (Table 8) shows all three effects, not only modality, are significantly smaller on IMED, and that modality is simply the effect whose CK+ margin was narrow

enough for shared attenuation to also cost it significance – visible only because both datasets shared an identical protocol. This addresses RQ1–RQ3: landmark pruning and resampling benefits generalize across the two datasets tested, while a fusion benefit does not, at the current sample size, extend from the larger to the smaller dataset with statistical confidence, and this difference is attributable to effect-size attenuation rather than a qualitatively distinct failure. As stated in the abstract, the IMED BorderlineSMOTE finding specifically is correction-method-dependent (significant under FDR, not under Bonferroni; Table 7) and should be carried forward with that qualification rather than as an unconditional conclusion.

Future work follows directly from Section 6.5: extending the imbalance comparison to cost-sensitive losses and other resampling variants under the same protocol; systematic hyperparameter tuning via nested cross-validation; evaluation protocols designed for subject-scarce data, such as leave-one-subject-out with seed-ensembling, compared directly against the GroupShuffleSplit-based protocol used here, since low statistical power on a 15-subject dataset is a property of the protocol's data efficiency as much as of the dataset; direct re-implementation of at least one prior CK+ baseline under this study's protocol and class set; and extension to in-the-wild video, IMED's mixed-emotion categories, or other datasets, to test whether the dataset-scale sensitivity found here for modality fusion is a general property of these three design choices. Further directions raised during review, and not addressed with new experiments in this revision, are named individually rather than folded into the list above: an architecture-capacity ablation, evaluating higher-capacity or attention-based models such as Vision Transformers under this study's own leakage-free protocol, to separate a protocol effect from an architecture effect in the CK+ literature gap; logging and reporting the BorderlineSMOTE random-oversampling fallback rate per fold; an ablation over frame-sampling strategies (uniform, motion-peak, fixed early-middle-late) in place of the fixed, late-sequence-biased sampling used here; alternative modality-fusion mechanisms (attention, gated fusion) beyond concatenation; an alternative, non-motion-magnitude landmark-importance score, such as class separability or classifier-weight-based scoring; and cross-dataset training or fine-tuning between CK+ and IMED to test generalization directly rather than only within-dataset subject-disjoint generalization.

Data and Code Availability

CK+ is available from its original publishers upon request [37]. IMED is available from its original publishers upon request [15]. Code implementing the pipeline and statistical testing will be made available from the corresponding author upon reasonable request, per the author team's decision. The full 30-by-30 Nemenyi pairwise p-value matrices underlying Section 5 are available from the corresponding author; Appendix 7 reproduces the subset of each matrix directly underlying the claims in this paper.

Acknowledgement

This research was supported by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia through the BIMA Fundamental Research Grant (Penelitian Fundamental Reguler, Multiyear Program Year 2), under Master Contract No. 133/C3/DT.05.00/PL-MULTITAHUN LANJUTAN/2026 and Derivative Contracts No. 16/LL6/AL.04.03/PL-MULTITAHUN LANJUTAN/2026 and No. 131/F.09/UDN-09/III/2026

Appendix Material

S1: Nemenyi pairwise comparisons against each dataset's best configuration

Tables 10 and 11 reproduce the row of each dataset's 30-by-30 Nemenyi matrix corresponding to comparisons against the best configuration (CK+: top30_combined_smoteTrue; IMED: top50_combined_smoteFalse), the specific numbers underlying the claim that 11 of 29 CK+ configurations are significantly worse than the best, and that zero of 29 IMED configurations differ significantly from the best, both at $p \leq 0.05$. Rows are sorted

by ascending p-value; pairwise cells not involving the best configuration are omitted here and available from the corresponding author.

Table 10. CK+: Nemenyi p-values, all 29 configurations versus top30_combined_smoteTrue.

Configuration	Nemenyi p
top100_xy_smoteFalse	1.77×10^{-9}
top100_optical_flow_smoteFalse	4.69×10^{-7}
top100_xy_smoteTrue	7.52×10^{-7}
top50_xy_smoteFalse	8.79×10^{-7}
top40_xy_smoteFalse	2.65×10^{-5}
top100_combined_smoteFalse	2.65×10^{-5}
top40_optical_flow_smoteFalse	6.08×10^{-5}
top50_optical_flow_smoteFalse	0.000229
top30_xy_smoteFalse	0.000381
top100_optical_flow_smoteTrue	0.001456
top30_optical_flow_smoteFalse	0.004060
– *significance threshold* $\alpha = 0.05$ –	
top50_xy_smoteTrue / top20_xy_smoteFalse	0.057250
top40_optical_flow_smoteTrue	0.067681
top50_optical_flow_smoteTrue	0.079664
top20_optical_flow_smoteFalse	0.167997
top50_combined_smoteFalse	0.529391
top40_xy_smoteTrue	0.571754
top100_combined_smoteTrue	0.634810
top30_optical_flow_smoteTrue	0.911730
top30_xy_smoteTrue	0.964876
top20_xy_smoteTrue	0.991143
top20_optical_flow_smoteTrue	0.997608
top40_combined_smoteFalse	0.998880
top20/30_combined_smoteFalse	0.999985
top50_combined_smoteTrue	0.999999
top40/20_combined_smoteTrue	1.000000

Table 11. IMED: Nemenyi p-values, all 29 configurations versus top50_combined_smoteFalse.

Configuration	Nemenyi p
top100_optical_flow_smoteTrue	0.062281
top100_optical_flow_smoteFalse	0.093351
top20_optical_flow_smoteFalse	0.655448
top50_optical_flow_smoteFalse	0.734451
top100_xy_smoteFalse	0.877640
top40_xy_smoteFalse	0.946321
top100_combined_smoteFalse	0.964876
top20_xy_smoteFalse	0.969825
top30_xy_smoteFalse / top50_optical_flow_smoteTrue	0.981498
top30_optical_flow_smoteFalse	0.996970
top100_xy_smoteTrue	0.998127
top40_optical_flow_smoteFalse	0.999515
top30_xy_smoteTrue	0.999977
top50_combined_smoteTrue	0.999993
top40_xy_smoteTrue	0.999996
top100_combined_smoteTrue	0.999999
remaining 13 configurations	1.000000

No IMED configuration falls below $\alpha = 0.05$; the closest is top100_optical_flow_smoteTrue at $p = 0.062$, itself 100% retention without resampling in single-modality form, the combination identified as generally weakest.

S2: MediaPipe Face Mesh landmark index reference

Figure 6 plots selection frequency against raw MediaPipe Face Mesh landmark index without naming facial regions. Figure 7 (Section 5.1) plots the same selection-frequency data directly at each index's canonical anatomical position, and is the primary reference for readers wanting a spatial view; the table below remains available as an index-level lookup for readers who want the underlying region-to-index mapping. Table 12 lists the landmark indices belonging to each named face-mesh contour region, extracted from the official `face_mesh_connections.py` source distributed with MediaPipe (Google, Apache License 2.0). These seven regions cover 152 of 468 landmarks; remaining landmarks belong to the dense facial tessellation without a single officially named group. "Left"/"right" follow MediaPipe's subject-perspective convention.

Table 12. MediaPipe Face Mesh named region index sets (468-point model).

Region	Count	Landmark indices
Lips	40	0, 13, 14, 17, 37, 39, 40, 61, 78, 80, 81, 82, 84, 87, 88, 91, 95, 146, 178, 181, 185, 191, 267, 269, 270, 291, 308, 310, 311, 312, 314, 317, 318, 321, 324, 375, 402, 405, 409, 415
Left eye	16	249, 263, 362, 373, 374, 380, 381, 382, 384, 385, 386, 387, 388, 390, 398, 466
Left eyebrow	10	276, 282, 283, 285, 293, 295, 296, 300, 334, 336
Right eye	16	7, 33, 133, 144, 145, 153, 154, 155, 157, 158, 159, 160, 161, 163, 173, 246
Right eyebrow	10	46, 52, 53, 55, 63, 65, 66, 70, 105, 107
Face oval (contour)	36	10, 21, 54, 58, 67, 93, 103, 109, 127, 132, 136, 148, 149, 150, 152, 162, 172, 176, 234, 251, 284, 288, 297, 323, 332, 338, 356, 361, 365, 377, 378, 379, 389, 397, 400, 454
Nose	24	1, 2, 4, 5, 6, 19, 45, 48, 64, 94, 97, 98, 115, 168, 195, 197, 220, 275, 278, 294, 326, 327, 344, 440

REFERENCES

[1] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[2] Xiangwei Mou, Yongfu Song, Xiuping Xie, Mingxuan You, and Rijun Wang. Res-RBG Facial Expression Recognition in Image Sequences Based on Dual Neural Networks. *Sensors*, 25(12):3829, June 2025.

[3] Xiaoliang Zhu, Zili He, Liang Zhao, Zhicheng Dai, and Qiaolai Yang. A Cascade Attention Based Facial Expression Recognition Network by Fusing Multi-Scale Spatio-Temporal Features. *Sensors*, 22(4):1350, February 2022.

[4] Chang Liu, Kaoru Hirota, Junjie Ma, Zhiyang Jia, and Yaping Dai. Facial Expression Recognition Using Hybrid Features of Pixel and Geometry. *IEEE Access*, 9:18876–18889, 2021.

[5] Chujie Xu, Wenjie Zheng, Yong Du, Tiejun Li, and Zhansheng Yuan. Online facial expression recognition based on graph convolution and long short memory networks. *Internet Technology Letters*, 8(1):e415, January 2025.

[6] Wenguo Li, Hongjun Lin, Yuyang Yan, and Zeqian Feng. Micro-expression recognition based on the fusion of facial landmark points and optical flow features. *Journal of Electronic Imaging*, 34(01), February 2025.

[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[8] Qingdu Li, Keting Fu, Jian Liu, Yishan Li, Qinze Ren, Kang Xu, Junxiu Fu, Na Liu, and Ye Yuan. Optimizing Class Imbalance in Facial Expression Recognition Using Dynamic Intra-Class Clustering. *Biomimetics*, 10(5):296, May 2025.

[9] Jiahong Yang, Zhisheng Lv, Kai Kuang, Sen Yang, Liuming Xiao, and Qiang Tang. RASN: Using Attention and Sharing Affinity Features to Address Sample Imbalance in Facial Expression Recognition. *IEEE Access*, 10:103264–103274, 2022.

[10] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing (ICIC 2005)*, pages 878–887. Springer, 2005.

[11] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[12] Ekin Yagis, Selamawet Workalemahu Atnafu, Alba García Seco De Herrera, Chiara Marzi, Riccardo Scheda, Marco Giannelli, Carlo Tessa, Luca Citi, and Stefano Diciotti. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific Reports*, 11(1):22544, November 2021.

[13] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, September 2023.

[14] Shan Li and Weihong Deng. A Deeper Look at Facial Expression Dataset Bias. *IEEE Transactions on Affective Computing*, 13(2):881–893, April 2022.

[15] Dewi Yanti Liliana, T. Basaruddin, and Imelda Ika Dian Oriza. The indonesian mixed emotion dataset (IMED): A facial expression dataset for mixed emotion recognition. In *Proceedings of the 2018 International Conference on Advances in Image Processing (AIVR 2018)*, pages 56–60. ACM, 2018.

[16] Sri Winarno, Farrikh Alzami, Dewi Agustini Santoso, Muhammad Naufal, Harun Al Azies, Rivaldo Mersis Brilianto, and Kalaiarasi Sonai Muthu. Efficient GRU-based facial expression recognition with adaptive loss selection. *Statistics, Optimization and Information Computing*, 14:3468–3499, 2025.

[17] Farid Ayeche and Adel Alti. Facial Expressions Recognition Based on Delaunay Triangulation of Landmark and Machine Learning. *Traitement du Signal*, 38(6):1575–1586, December 2021.

[18] Omer Faruk Soylemez and Burhan Ergen. Facial Landmark Based Region of Interest Localization for Deep Facial Expression Recognition. *Tehnicki vjesnik - Technical Gazette*, 29(1), February 2022.

[19] Lei Liao, Yu Zhu, Bingbing Zheng, Xiaoben Jiang, and Jiajun Lin. FERGCN: facial expression recognition based on graph convolution network. *Machine Vision and Applications*, 33(3):40, May 2022.

[20] Mojtaba Kolahdouzi, Alireza Sepas-Moghaddam, and Ali Etemad. FaceTopoNet: Facial Expression Recognition Using Face Topology Learning. *IEEE Transactions on Artificial Intelligence*, 4(6):1526–1539, December 2023.

[21] Chenghao Fu, Wenzhong Yang, Danny Chen, and Fuyuan Wei. AM3F-FlowNet: Attention-Based Multi-Scale Multi-Branch Flow Network. *Entropy*, 25(7):1064, July 2023.

[22] Chao Wan, Wenbing Zhang, Yadong Chen, Liangliang Song, and Peng Cheng. ARTNet for Micro-Expression Recognition. *Sensors*, 26(1):247, December 2025.

[23] Navneet Gupta, R. Vishnu Priya, and Chandan Kumar Verma. ERFN: Leveraging context for enhanced emotion detection. *International Journal of Advanced Computer Science and Applications*, 15(6):617–632, 2024.

[24] Ke Zhao, Xuanyu Liu, and Guangqian Yang. M3ENet: A Multi-Modal Fusion Network for Efficient Micro-Expression Recognition. *Sensors*, 25(20):6276, October 2025.

[25] Ahmed Mostafa, Hala El-Sayed, and Mohamed Belal. Facial Expressions Recognition Via CNNCraft-net for Static RGB Images. *International Journal of Intelligent Engineering and Systems*, 14(4):410–421, August 2021.

[26] Jie Shao and Yanyang Shi. Dynamic data augmentation parameter learning-based unbalanced facial emotion recognition. *Journal of Electronic Imaging*, 31(05), September 2022.

[27] Damien Dablain, Bartosz Krawczyk, and Nitesh V. Chawla. DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):6390–6404, September 2023.

[28] Javad Hassannataj Joloudari, Abdolreza Marefat, Mohammad Ali Nematollahi, Solomon Sunday Oyelere, and Sadiq Hussain. Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. *Applied Sciences*, 13(6):4006, March 2023.

[29] Muhammad Mujahid, Erol Kına, Furqan Rustam, Monica Gracia Villar, Eduardo Silva Alvarado, Isabel De La Torre Diez, and Imran Ashraf. Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. *Journal of Big Data*, 11(1):87, June 2024.

[30] Gilles Vandewiele, Isabelle Dehaene, György Kovács, Lucas Sterckx, Olivier Janssens, Femke Ongenae, Femke De Backere, Filip De Turck, Kristien Roelens, Johan Decruyenaere, Sofie Van Hoecke, and Thomas Demeester. Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. *Artificial Intelligence in Medicine*, 111:101987, January 2021.

[31] Matthew Rosenblatt, Link Tejavibulya, Rongtao Jiang, Stephanie Noble, and Dustin Scheinost. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, 15(1):1829, February 2024.

[32] M. Kas, Y. El Merabet, Y. Ruichek, and R. Messoussi. New framework for person-independent facial expression recognition combining textural and shape analysis through new feature extraction approach. *Information Sciences*, 549:200–220, March 2021.

[33] Lihua Bi, Shenbo Tang, and Canlin Li. A facial expression recognition method based on improved VGG19 model. *International Journal of Advanced Computer Science and Applications*, 15(7):249–256, 2024.

[34] A. Shanimol and J. Charles. ResNet50 and GRU: A synergistic model for accurate facial emotion recognition. *International Journal of Advanced Computer Science and Applications*, 15(8):611–620, 2024.

[35] Aditya Wikan Mahastama, Edwin Mahendra, Antonius Rachmat Chrismanto, Maria Nila Anggia Rini, and Andhika Galuh Prabawati. Facial expression classification system using stacked CNN. *International Journal of Advanced Computer Science and Applications*, 15(10):464–472, 2024.

[36] Limas Jaya Akeh and Gede Putra Kusuma. Mixed emotion recognition through facial expression using transformer-based model. *Statistics, Optimization and Information Computing*, 13(2):531–546, 2025.

[37] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW)*, pages 94–101, 2010.

[38] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[39] Huafei Xiao, Wenbo Li, Guanzhong Zeng, Yingzhang Wu, Jiyong Xue, Juncheng Zhang, Chengmou Li, and Gang Guo. On-Road Driver Emotion Recognition Using Facial Expression. *Applied Sciences*, 12(2):807, January 2022.