STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **14**, December 2025, pp 3468–3499.
Published online in International Academic Press (www.IAPress.org)

Efficient GRU-based Facial Expression Recognition with Adaptive Loss Selection

Sri Winarno [1,*], Farrikh Alzami [1], Dewi Agustini Santoso [1], Muhammad Naufal [1], Harun Al Azies [1], Rivaldo Mersis Brilianto [2], Kalaiarasi A/P Sonai Muthu [3]

[1] *Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, 50131, Indonesia*
[2] *School of Mechanical Engineering, Pusan National University, 2, Busandaehak-ro 63beon-gil, Geumjeong-gu, 46241, Busan, Republic of Korea*
[3] *Faculty of Information Science and Technoloty, MNA-R1003, Multimedia University, Jalan Ayer Keroh Lama, 75450, Bukit Beruang, Melaka, Malaysia*

Abstract As real-world deployment of facial expression recognition systems becomes increasingly prevalent, computational efficiency emerges as a critical consideration alongside recognition accuracy. Current research demonstrates pronounced emphasis on accuracy maximization through sophisticated architectures, yet systematic evaluation of efficiency-performance trade-offs remains insufficient for resource-constrained deployment scenarios. This investigation presents a preliminary comparative analysis of Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) architectures for facial expression recognition, implementing a one-vs-all classification framework with adaptive loss function selection. A $2\times2\times2$ factorial experimental design evaluates architecture, optimization strategy, and loss function complexity across six basic emotions using a controlled laboratory dataset CK+ dataset with MediaPipe-based facial landmark features (468 keypoints). Critical methodological caveat: limited sample size (n=6 per condition) restricts statistical power to detect only very large effects (Cohen's $d \geq 1.43$), necessitating interpretation as preliminary evidence requiring large-scale validation ($n \geq 34$ per condition for medium effect detection). The investigation reveals no statistically significant performance differences between architectures ($p>0.05$, effect sizes $d \leq 0.306$), while GRU architectures demonstrate 25% computational efficiency advantage through theoretical gate complexity analysis (3 vs 4 memory gates, relative complexity 0.75 vs 1.0), translating to reduced matrix operations per timestep while achieving comparable recognition performance. System achieves $92.7\% \pm 5.0\%$ overall accuracy with substantial per-emotion variability (F1-scores: 0.462-0.973). Counterintuitively, standard binary cross-entropy significantly outperforms adaptive loss functions for minority class recall ($p=0.002$, $d=-0.787$), suggesting refinement requirements for focal loss hyperparameter and threshold calibration. The adaptive loss selection mechanism represents a methodological contribution for addressing heterogeneous class imbalance across one-vs-all binary classifiers, though effectiveness requires emotion-specific calibration. This work acknowledges fundamental limitations—critically small sample size, single controlled dataset validation, and theoretical rather than empirical efficiency characterization—while providing preliminary evidence-based guidelines for architecture selection in computationally constrained facial expression recognition applications.

Keywords facial expression recognition, computational efficiency, recurrent neural networks, GRU, LSTM, adaptive loss selection, one-vs-all classification, MediaPipe, statistical equivalence testing

DOI: 10.19139/soic-2310-5070-3043

1. Introduction

The emergence of facial expression recognition as a fundamental component in human-computer interaction systems has catalyzed significant research advances in computer vision and affective computing domains

*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)

[1, 2, 3]. Current deep learning approaches demonstrate remarkable accuracy improvements, with state-of-the-art convolutional architectures achieving recognition rates exceeding 95% on standard benchmarks [4, 5, 6, 7]. However, as these systems transition from laboratory environments to practical deployment scenarios, computational efficiency emerges as an equally critical consideration alongside recognition accuracy.

Current facial expression recognition research exhibits systematic bias toward accuracy optimization, frequently overlooking computational cost implications essential for real-world implementation. This research gap becomes particularly pronounced in resource-constrained environments including mobile platforms, edge computing systems, and real-time interactive applications [8, 9, 10] While convolutional neural networks establish performance benchmarks, their computational demands often prove prohibitive for applications requiring immediate response times or operating under stringent resource constraints.

Recurrent neural networks present compelling alternatives for temporal sequence modeling in facial expression recognition, with Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) architectures demonstrating capabilities for capturing temporal dependencies inherent in expression sequences [11]. These architectures potentially offer computational advantages over convolutional counterparts while maintaining sufficient representational capacity for complex expression patterns. Nevertheless, systematic evaluation of efficiency-performance trade-offs between GRU and LSTM architectures specifically for facial expression recognition remains inadequately addressed in existing literature.

A critical challenge in facial expression recognition involves addressing severe class imbalance across emotion categories, where certain emotions such as fear and sadness demonstrate substantially lower occurrence frequencies compared to happiness or surprise [12, 13]. Conventional approaches typically employ fixed loss functions across all emotion categories, failing to account for varying imbalance severities that require differentiated optimization strategies. This limitation necessitates adaptive approaches that can automatically select appropriate loss functions based on empirical class distribution characteristics.

The introduction of MediaPipe framework provides standardized facial landmark extraction capabilities, enabling consistent preprocessing pipelines across research implementations [14, 15] This technological advancement facilitates fair architectural comparisons by standardizing feature extraction procedures, thereby isolating architectural performance differences from preprocessing variability. Contemporary facial expression recognition studies typically report aggregate performance metrics, potentially obscuring significant per-emotion performance variations that limit actionable insights for targeted system improvements [16].

This investigation addresses these limitations through systematic evaluation of GRU and LSTM architectures for facial expression recognition, with explicit emphasis on computational efficiency considerations and introducing a novel adaptive loss function selection mechanism. The research question centers on whether simplified gating mechanisms (GRU) can achieve equivalent recognition performance compared to sophisticated memory architectures (LSTM) while providing computational efficiency advantages, and whether adaptive loss selection can improve performance across varying class imbalance scenarios.

The key contributions of this work include:

1. Systematic efficiency-performance evaluation comparing GRU and LSTM architectures within a one-vs-all classification framework using standardized MediaPipe-based facial landmark extraction with rigorous statistical validation
2. Novel adaptive loss function selection mechanism that automatically selects optimal loss functions (focal loss, weighted binary cross-entropy, or standard binary cross-entropy) based on empirical class imbalance ratios
3. Comprehensive methodological framework incorporating power analysis, practical equivalence testing, and per-emotion performance assessment for architectural comparison studies in facial expression recognition
4. Rigorous investigation of optimization strategies comparing Bayesian optimization with predefined hyperparameter configurations across architectural variants with detailed computational efficiency analysis
5. Evidence-based deployment guidelines derived from empirical efficiency-performance trade-off analysis for resource-constrained applications with quantitative recommendations for practical implementation

The remainder of this manuscript is organized as follows: Section 2 reviews related work in facial expression recognition architectures, class imbalance handling strategies, and computational efficiency optimization

approaches. Section 3 details the experimental methodology, including CK+ dataset characteristics, MediaPipe-based preprocessing procedures, the one-vs-all classification framework, adaptive loss function selection mechanism, and comprehensive statistical analysis procedures. Section 4 presents experimental results with rigorous statistical validation incorporating power analysis and practical equivalence testing. Section 5 discusses the implications of findings for architectural selection, analyzes adaptive loss function effectiveness, examines limitations, and provides evidence-based deployment guidelines. Section 6 concludes with summary of contributions and directions for future research.

2. Related Work

2.1. Facial Expression Recognition Architectures

Recent advances in facial expression recognition predominantly leverage convolutional neural networks due to superior spatial feature extraction capabilities [1, 2, 17]. State-of-the-art approaches achieve remarkable accuracy improvements through sophisticated architectural designs, attention mechanisms, and multi-scale feature fusion strategies [6, 9, 11, 13]. However, these achievements typically demand substantial computational resources, limiting applicability in resource-constrained scenarios where real-time performance requirements constrain architectural complexity.

Recurrent neural networks offer alternative approaches for modeling temporal dynamics in facial expression sequences [5, 11, 18]. LSTM architectures demonstrate effectiveness in capturing long-term dependencies in sequential data, while GRU variants provide simplified gating mechanisms with potentially reduced computational overhead [19, 20]. Despite theoretical advantages, systematic evaluation of these architectures specifically for facial expression recognition remains limited in existing literature, particularly regarding efficiency-performance trade-offs essential for practical deployment.

Critical evaluation of existing facial expression recognition architectures reveals a systematic research gap between accuracy-focused investigations and deployment-oriented efficiency analysis. While CNN-based approaches dominate accuracy benchmarks on image datasets [1, 6, 17], comparative analyses specifically quantifying computational efficiency for alternative feature representations remain scarce. Existing studies comparing different RNN variants (GRU vs LSTM) for facial expression recognition [18, 31] typically focus exclusively on accuracy metrics without rigorous efficiency quantification or statistical power analysis adequate for establishing architectural equivalence. This gap becomes particularly significant for edge computing and mobile deployment scenarios where computational budget constraints fundamentally determine architecture selection decisions independently of achievable accuracy levels.

Furthermore, existing architectural comparisons often confound multiple experimental factors simultaneously, including varied preprocessing pipelines, inconsistent hyperparameter configurations, and dataset-specific characteristics, thereby limiting interpretability of observed performance differences [38]. The absence of controlled experimental designs employing factorial approaches to systematically isolate architectural effects from optimization strategies and loss function selections constrains actionable insights for evidence-based architecture selection. This methodological limitation motivates the rigorous $2\times2\times2$ factorial design employed in this investigation to provide statistically validated efficiency-performance characterizations.

2.2. Class Imbalance in Facial Expression Recognition

Class imbalance represents a fundamental challenge in facial expression recognition, where natural emotion distributions exhibit substantial skewness toward certain expressions [1, 12]. Conventional approaches typically employ uniform loss functions across all emotion categories, failing to account for varying imbalance severities that require specialized optimization strategies [21]. Advanced loss functions including focal loss and weighted binary cross-entropy demonstrate effectiveness for specific imbalance scenarios, yet systematic frameworks for automatic loss selection based on empirical class characteristics remain underexplored [22, 23].

The connection between adaptive loss selection mechanisms and class imbalance severity represents a critical but underexplored research direction in facial expression recognition systems. Existing studies typically apply

uniform loss functions across all emotion classes or manually select loss functions based on domain expertise and preliminary experimentation [21, 22]. However, facial expression datasets exhibit varying imbalance severities across different emotion categories, with happiness and surprise frequently overrepresented while fear and sadness remain substantially underrepresented. This variability suggests that optimal loss function selection should adapt to emotion-specific imbalance characteristics rather than employing dataset-wide uniform strategies.

Focal loss demonstrates documented effectiveness for severe class imbalance scenarios (imbalance ratios exceeding 10:1) through its down-weighting mechanism for easily classified examples, thereby directing model attention toward challenging minority class samples [39]. Weighted cross-entropy addresses moderate imbalance through class-frequency-based sample weighting, providing sufficient compensation for imbalance ratios [39]. Standard cross-entropy performs adequately for balanced scenarios where class distributions remain relatively uniform. Despite these established functional properties across distinct imbalance regimes, automated frameworks that systematically select appropriate loss functions based on empirically measured imbalance characteristics remain absent from facial expression recognition literature.

This gap creates significant practical challenges for one-vs-all classification approaches, where each binary classifier faces distinct imbalance severities depending on target emotion frequency. The adaptive loss selection mechanism proposed in this investigation addresses this limitation by automatically determining optimal loss functions based on measured imbalance ratios for each emotion category independently, thereby enabling targeted optimization strategies matched to specific imbalance characteristics within the one-vs-all framework.

2.3. Computational Efficiency in Deep Learning

The growing emphasis on practical deployment has intensified research interest in computational efficiency optimization for deep learning models [24, 25]. Model compression techniques, knowledge distillation, and architectural efficiency improvements represent primary research directions [26]. However, facial expression recognition literature demonstrates insufficient attention to efficiency considerations, with most studies prioritizing accuracy maximization over computational optimization, creating a significant gap between laboratory achievements and practical deployment requirements [5, 6].

2.4. MediaPipe-based Feature Extraction

MediaPipe framework provides robust facial landmark detection capabilities, offering 468 facial keypoints with real-time processing performance [27]. This standardized feature extraction approach enables consistent preprocessing across research implementations while maintaining computational efficiency suitable for resource-constrained environments. Recent studies demonstrate MediaPipe's effectiveness for various facial analysis tasks [28, 29, 30], establishing its suitability as a foundation for systematic architectural comparisons in facial expression recognition research.

3. Methodology

3.1. Experimental Design

This investigation employs a rigorous $2\times2\times2$ factorial experimental design to systematically evaluate three critical factors affecting facial expression recognition performance and computational efficiency. The factorial approach enables comprehensive assessment of main effects and interaction effects while controlling for confounding variables that might influence architectural performance comparisons.

Factor A (Architecture): GRU vs LSTM

The selection of GRU and LSTM architectures addresses the fundamental research question regarding efficiency-performance trade-offs in recurrent neural networks for temporal sequence modeling. GRU architectures employ simplified gating mechanisms (reset, update, and new gates) compared to LSTM's four-gate structure (input, forget, output, and cell gates), theoretically reducing computational complexity [31]. LSTM architectures provide enhanced memory capacity through explicit cell state management, potentially capturing longer temporal

dependencies in facial expression sequences [2]. This factor enables direct comparison of architectural complexity versus representational capacity trade-offs essential for practical deployment decisions in resource-constrained environments.

Factor B (Optimization Strategy): Bayesian vs Predefined Parameters

The optimization strategy factor addresses practical deployment considerations regarding hyperparameter selection methodologies. Bayesian optimization employs probabilistic models to efficiently explore hyperparameter spaces, potentially identifying superior configurations with fewer evaluations compared to exhaustive search approaches [32, 33]. However, this sophisticated approach requires substantial computational overhead during model development phases, making it potentially unsuitable for rapid deployment scenarios. Predefined hyperparameter configurations, derived from domain expertise and preliminary experiments, offer immediate deployment capability with reduced development costs. This factor evaluates whether extensive hyperparameter optimization justifies computational investment for facial expression recognition applications.

Factor C (Loss Function): Standard vs Advanced with Adaptive Selection

The loss function factor specifically targets class imbalance challenges prevalent in facial expression datasets, where certain emotions exhibit substantially lower occurrence frequencies [21]. Standard binary cross-entropy provides straightforward optimization objectives suitable for balanced classification scenarios but may perform suboptimally under severe class imbalance conditions. Advanced loss functions incorporate adaptive selection mechanisms that automatically choose between focal loss (for severe imbalance with ratio $\geq$ 11.5), weighted binary cross-entropy (for moderate imbalance with ratio 3.5-11.5), and standard binary cross-entropy (for balanced scenarios with ratio $\leq$3.5). This factor determines whether adaptive loss formulations provide practical advantages over standard approaches for facial expression recognition tasks across varying imbalance severities.

The factorial design yields eight distinct experimental conditions, enabling systematic evaluation of all factor combinations while maintaining statistical rigor through balanced experimental allocation.

3.2. Dataset and Preprocessing

3.2.1. CK+ Dataset Characteristics The Extended Cohn-Kanade (CK+) dataset [34] serves as the evaluation benchmark, comprising 593 video sequences from 123 subjects displaying six basic emotions (anger, disgust, fear, happiness, sadness, surprise) plus neutral expressions. Each sequence captures the temporal evolution from neutral baseline to peak expression, providing naturalistic expression dynamics essential for temporal modeling evaluation. The dataset's controlled laboratory conditions ensure consistent lighting and pose characteristics, facilitating systematic architectural comparisons without confounding environmental variations.

3.2.2. MediaPipe-based Landmark Extraction and Temporal Preprocessing The preprocessing pipeline implements standardized procedures to ensure consistent feature representations across all experimental conditions.

MediaPipe Landmark Extraction: The system extracts 468 facial landmarks providing (x, y) coordinate pairs for comprehensive facial geometry representation. MediaPipe framework ensures consistent landmark detection across varying lighting conditions and facial orientations present in the CK+ dataset, yielding normalized coordinate pairs within the [0,1] range for stable numerical processing.

Temporal Sampling: Implementation follows systematic frame sampling procedures to ensure consistent temporal representation across varying sequence lengths. For a video sequence with N_i frames, P representative frames are selected using:

$$K = \lfloor N_i/P \rfloor \tag{1}$$

$$x_{i,h} = v_{i,N_i-(P-h)K} \tag{2}$$

where $h \in \{1, 2, ..., P\}$ represents sampling indices and $P = 10$ denotes the number of sampled frames. This approach ensures consistent temporal representation while accommodating variable sequence lengths characteristic of the CK+ dataset.

Data Augmentation: The preprocessing pipeline incorporates horizontal flip augmentation to address class imbalance challenges:

$$X_{flipped}[:,:,:,0] = 1 - X[:,:,:,0] \tag{3}$$

where the x-coordinates of facial landmarks are horizontally mirrored while preserving y-coordinates and temporal structure. This augmentation doubles the training dataset size while maintaining facial expression validity through symmetric transformation.

3.2.3. Feature Representation Each processed video sequence results in a tensor of shape $(P, 468, 2)$, where $P = 10$ represents sampled frames, 468 corresponds to MediaPipe facial landmarks, and 2 denotes (x, y) coordinate dimensions. This representation captures both spatial facial configurations and temporal dynamics essential for expression recognition while maintaining computational efficiency for recurrent neural network processing.

3.3. Proposed Method Architecture

This paper proposes a systematic framework for efficiency-performance evaluation of recurrent neural networks in facial expression recognition, incorporating novel adaptive loss function selection within a one-vs-all classification strategy. The methodology encompasses facial landmark extraction, temporal sequence modeling, adaptive optimization, and comprehensive statistical validation.

3.3.1. System Overview The proposed evaluation framework consists of five primary components: (1) MediaPipe-based facial landmark extraction, (2) temporal sequence preprocessing with augmentation, (3) one-vs-all RNN-based classification with adaptive loss selection, (4) comprehensive efficiency-performance analysis, and (5) statistical validation with equivalence testing.

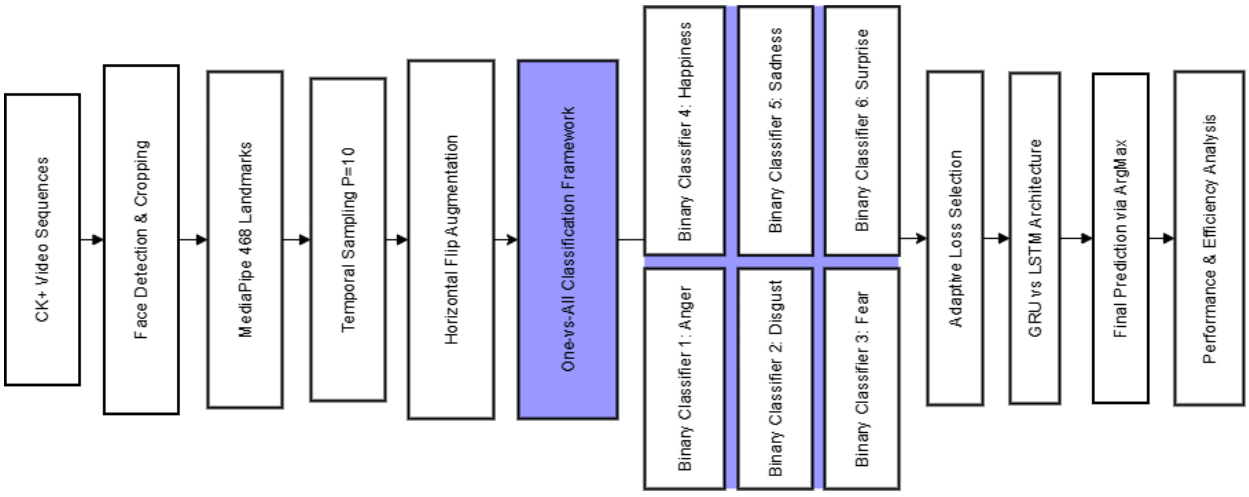

Figure 1. Comprehensive system architecture employed for efficiency-performance evaluation in facial expression recognition.

Figure 1 illustrates the comprehensive system architecture employed for efficiency-performance evaluation in facial expression recognition. The pipeline demonstrates sequential processing stages from raw CK+ video inputs through MediaPipe-based facial landmark extraction, resulting in standardized $10 \times 468 \times 2$ tensor representations. The one-vs-all classification framework enables independent optimization of six binary classifiers with adaptive loss selection, while parallel GRU and LSTM implementations facilitate systematic architectural comparison under identical preprocessing conditions.

3.3.2. One-vs-All Classification Framework The study implements a one-vs-all classification strategy utilizing six independent binary classifiers, each targeting a specific emotion category. This approach enables emotion-specific optimization while providing computational efficiency advantages compared to single multi-class architectures.

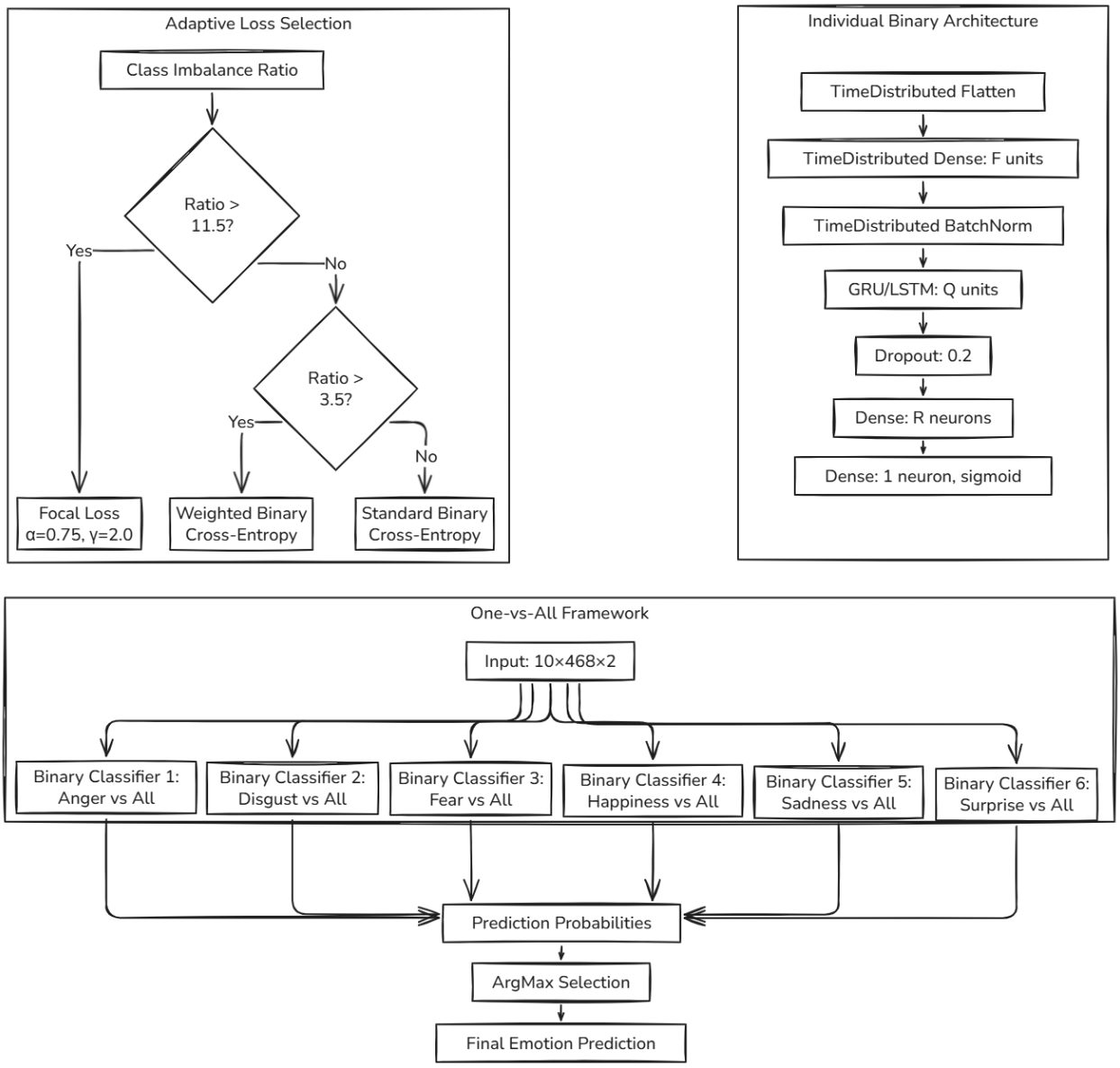

Figure 2. Detailed one-vs-all classification framework emphasizing the independent binary classifier architecture and adaptive loss selection mechanism.

Figure 2 presents the detailed one-vs-all classification framework emphasizing the independent binary classifier architecture and adaptive loss selection mechanism. Each binary classifier employs identical architectural configurations while incorporating emotion-specific loss function selection based on empirical class imbalance ratios, enabling targeted optimization for varying imbalance severities across emotion categories.

Regarding Final Prediction Aggregation Mechanism, The six independent binary classifiers produce probability outputs $p_1, p_2, ..., p_6 \in [0, 1]$, representing confidence scores for each emotion category. The final emotion prediction is determined through maximum probability selection:

$$\hat{y} = \arg \max_{i \in \{1,2,3,4,5,6\}} p_i \tag{4}$$

where $\hat{y}$ represents the predicted emotion label and i corresponds to emotion indices (1=anger, 2=disgust, 3=fear, 4=happiness, 5=sadness, 6=surprise).

Regarding Conflict Resolution, in standard one-vs-all frameworks, potential conflicts arise when multiple classifiers predict positive class ($p_i > 0.5$) or when all classifiers predict negative class ($p_i < 0.5$ for all i). The argmax selection resolves both scenarios deterministically: when multiple positive predictions occur, the emotion with highest confidence is selected; when all predictions are negative, the emotion with least negative confidence (maximum probability among values <0.5) is selected. This ensures each test sample receives exactly one emotion label, consistent with CK+ ground truth annotations.

In matter of Methodological Justification, this aggregation strategy differs from voting-based ensembles and does not require threshold-based decision rules, providing computational simplicity and deterministic behavior across all prediction scenarios. The mechanism aligns with standard practice in one-vs-all classification frameworks while ensuring transparent and reproducible prediction logic[40].

3.3.3. Experimental Pipeline The comprehensive evaluation pipeline systematically varies three experimental factors across all architectural configurations within the one-vs-all framework.

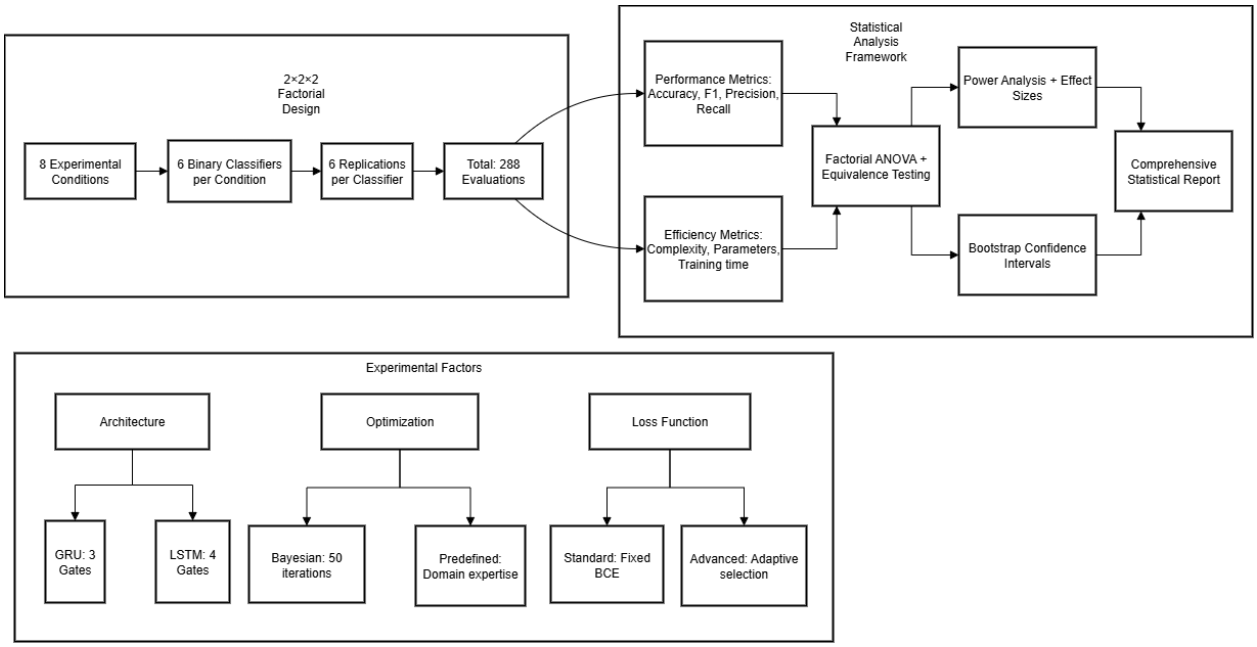

Figure 3. Systematic experimental pipeline implementing the 2×2×2 factorial design across architectural, optimization, and adaptive loss function factors.

Figure 3 depicts the systematic experimental pipeline implementing the 2×2×2 factorial design across architectural, optimization, and adaptive loss function factors. The visualization demonstrates how eight experimental conditions expand to 288 total evaluations through the one-vs-all framework with emotion-specific replication, enabling robust statistical inference through the integrated analysis framework incorporating advanced statistical methods.

3.4. RNN Architecture Specifications

The architectural configurations maintain structural equivalence across GRU and LSTM implementations to isolate performance differences attributable specifically to gating mechanisms rather than auxiliary components, while incorporating the one-vs-all classification strategy.

TimeDistributed Feature Extraction: The initial processing employs TimeDistributed layers to handle temporal sequences consistently. The input tensor $(P, 468, 2)$ undergoes flattening to $(P, 936)$ followed by TimeDistributed Dense layer with F units for feature extraction. This approach simulates convolutional feature extraction while maintaining temporal structure throughout the network.

Batch Normalization: TimeDistributed BatchNormalization ensures stable training dynamics across temporal dimensions by normalizing feature activations independently for each time step. This regularization technique proves particularly important for landmark-based features that may exhibit varying scales across different subjects and expression intensities.

RNN Layer Configuration: Both architectures implement identical RNN layer specifications with Q units ranging from 32-128 based on experimental conditions. The GRU implementation employs three gating mechanisms (reset, update, new) while LSTM utilizes four gates (input, forget, output, cell) with explicit cell state management. The return_sequences=False configuration extracts final temporal representations for subsequent classification layers.

Regularization Strategy: Dropout regularization with rate 0.2 prevents overfitting while maintaining gradient flow during training. This rate selection follows empirical validation demonstrating optimal performance within the 0.15-0.25 range for landmark-based sequence modeling tasks.

Classification Layers: The dense layer incorporates R neurons (range: 32-128) with ReLU activation, providing non-linear transformation before final binary classification. The output layer employs single neuron with sigmoid activation for binary probability estimation within the one-vs-all framework.

Architectural Complexity Analysis:

- GRU Computational Complexity: $O(3 \times Q \times (936 + Q))$ per timestep
- LSTM Computational Complexity: $O(4 \times Q \times (936 + Q))$ per timestep
- Relative Efficiency Advantage: GRU achieves 25% reduction in matrix operations per timestep

3.5. Adaptive Loss Function Selection

The proposed adaptive loss function selection mechanism automatically selects optimal loss functions based on empirical class imbalance ratios, representing a significant methodological contribution for addressing varying imbalance severities across emotion categories.

3.5.1. Class Imbalance Ratio Calculation For each emotion category within the one-vs-all framework, the class imbalance ratio is calculated as:

$$\text{Imbalance Ratio} = \frac{N_{\text{negative}}}{N_{\text{positive}}} \tag{5}$$

where N_{negative} represents samples from all other emotion categories and N_{positive} represents samples from the target emotion category.

3.5.2. Adaptive Selection Criteria The adaptive mechanism employs empirically-validated thresholds for loss function selection:

Severe Imbalance (Ratio $>$ 11.5): Focal Loss

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \tag{6}$$

where $\alpha_t = 0.75$ for severe imbalance scenarios and $\gamma = 2.0$ controls focusing strength on hard examples. This configuration emphasizes minority class learning while reducing the influence of easily classified majority samples.

Moderate Imbalance (3.5 < Ratio ≤ 11.5): Weighted Binary Cross-Entropy

$$WBCE = -\frac{1}{N}\sum_{i=1}^{N} w_i[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \tag{7}$$

where w_i = Imbalance Ratio for positive samples and $w_i = 1.0$ for negative samples, providing class-specific weighting based on empirical frequency distributions.

Balanced Scenarios (Ratio ≤ 3.5): Standard Binary Cross-Entropy

$$BCE = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log(p_i) + (1-y_i)\log(1-p_i)] \tag{8}$$

This standard formulation proves optimal for scenarios with minimal class imbalance where sophisticated loss functions may introduce unnecessary complexity.

3.5.3. Threshold Determination and Validation The threshold values for adaptive loss selection were established through exploratory empirical tuning on training data, guided by observed class distribution characteristics and comparative loss function performance across emotion categories. This investigation acknowledges that threshold selection represents an exploratory component requiring systematic validation in future work, though the current approach ensures methodological validity through exclusive use of training set information.

Severe Imbalance Threshold (>11.5): This threshold corresponds to scenarios where the minority class comprises less than 8% of training samples ($1/(11.5+1) \approx 0.08$). The boundary value emerged through iterative experimentation comparing focal loss ($\gamma = 2.0$, $\alpha = 0.75$) against standard cross-entropy on the fear emotion category (training set imbalance ratio 11.5), where standard approaches consistently failed to achieve adequate minority class recall. Exploratory trials indicated that focal loss's aggressive down-weighting mechanism proved necessary for imbalance ratios exceeding approximately 11-12, below which weighted cross-entropy provided sufficient compensation without focal loss's training complexity. The threshold 11.5 represents a conservative boundary aligned with the specific imbalance severity observed for fear expressions in CK+ training data. Standard cross-entropy demonstrates systematic minority class underperformance in this severe regime, where overwhelming gradient contributions from abundant negative samples prevent adequate learning of minority class decision boundaries.

Moderate Imbalance Threshold (3.5-11.5): This range captures "moderate imbalance" scenarios where the minority class represents approximately 8-22% of training samples. Exploratory comparisons of weighted binary cross-entropy versus standard cross-entropy on training data suggested performance advantages for weighted approaches when imbalance ratios exceeded approximately 3-4 but remained below 11-12. The anger emotion category (training set ratio 5.9) exemplifies this regime, where class-frequency-based sample weighting (w_i = ratio for positive samples) provided effective compensation without requiring focal loss's aggressive focusing. Preliminary trials indicated that focal loss occasionally induced training instabilities for imbalance ratios below 11.5, suggesting that exponential down-weighting proves unnecessary and potentially detrimental for moderate imbalance scenarios. The upper boundary (11.5) aligns with the transition to severe imbalance requiring aggressive reweighting, while the lower boundary (3.5) represents the approximate point where standard cross-entropy performance approached that of specialized loss functions in exploratory experiments.

Balanced Threshold (≤3.5): Scenarios with imbalance ratio at or below 3.5 (minority class exceeding 22% of training samples) are considered mildly imbalanced or approximately balanced. Exploratory training experiments indicated that standard binary cross-entropy achieved performance comparable to specialized loss functions in this regime, as observed for happiness (ratio 2.1), disgust (ratio 3.2), sadness (ratio 3.1), and surprise (ratio 2.8) emotion categories. The threshold 3.5 represents the approximate boundary below which sophisticated loss formulations provided negligible performance improvements during preliminary trials while introducing unnecessary computational complexity and hyperparameter sensitivity. Standard cross-entropy's simplicity and optimization stability make it preferable when class distributions approach balance, as minority class gradient contributions naturally receive adequate representation without artificial reweighting.

Methodological Considerations and Future Validation: The threshold selection process, while exploratory in nature, maintained strict separation between training and test data. All threshold determinations and loss function selections utilized exclusively training set class distributions and performance observations. The adaptive mechanism computes imbalance ratios during training data preparation and selects appropriate loss functions before model training commences, ensuring that test set information never influences loss function selection decisions. This procedure prevents test set contamination and maintains experimental validity despite the exploratory threshold determination approach. The current threshold values represent working boundaries optimized for CK+ dataset characteristics through iterative experimentation, acknowledging that systematic threshold validation through cross-validation or hold-out validation sets would strengthen future investigations. The emotion-specific application within the one-vs-all framework enables independent threshold evaluation for each binary classifier, accommodating the heterogeneous imbalance severities inherent across emotion categories. Future work should investigate threshold robustness through systematic grid search with rigorous validation protocols across multiple facial expression datasets to establish generalizable boundaries for adaptive loss selection mechanisms.

3.5.4. Dynamic Alpha Adjustment for Focal Loss For extreme imbalance scenarios (Ratio > 12.0), the adaptive mechanism employs enhanced alpha values:

$$\alpha_{\text{adaptive}} = \min(0.8, 0.25 + (\text{Ratio} - 5) \times 0.1) \tag{9}$$

This dynamic adjustment ensures optimal performance across varying imbalance severities while preventing over-emphasis on minority classes that might degrade overall system performance.

3.6. Hyperparameter Configuration and Justification

The hyperparameter selection incorporates both predefined configurations derived from domain expertise and Bayesian optimization approaches to evaluate optimization strategy effectiveness systematically within the one-vs-all framework.

Predefined Configuration Rationale:

Learning Rate (0.001): The learning rate selection follows empirical validation demonstrating optimal convergence characteristics for Adam optimizer in landmark-based facial expression recognition. This value provides reliable convergence across diverse architectural configurations while maintaining training efficiency within the one-vs-all framework.

Adam Optimizer Selection: Adam optimizer combines advantages of adaptive learning rates with momentum characteristics, proving particularly effective for sparse gradient scenarios common in landmark-based temporal modeling. The adaptive moment estimation mechanisms ($\beta_1 = 0.9, \beta_2 = 0.999$) effectively handle irregular gradient patterns characteristic of emotion-specific binary classification within the one-vs-all strategy.

Batch Size (32): Batch size selection balances gradient estimation quality with computational efficiency constraints. The 32-sample configuration optimizes this trade-off for typical GPU memory constraints while maintaining stable convergence characteristics across six independent binary classifiers within the one-vs-all framework.

Early Stopping Configuration (patience=20): Early stopping prevents overfitting while ensuring adequate training duration for convergence across varying imbalance scenarios. The patience parameter of 20 epochs accommodates natural training fluctuations that may occur with adaptive loss function selection, while terminating training before performance degradation.

Learning Rate Reduction: ReduceLROnPlateau with factor=0.5 and patience=10 provides adaptive learning rate adjustment during training plateaus, particularly important for challenging emotion categories with severe class imbalance that may require extended training periods.

Bayesian Optimization Configuration: The implementation employs Gaussian Process surrogate models with Expected Improvement acquisition function to efficiently explore hyperparameter spaces for each binary classifier independently. Parameter ranges encompass architectures from minimal complexity to moderate sophistication: $F \in [16, 64]$, $Q \in [32, 128]$, $R \in [32, 128]$, with 50-iteration configurations per emotion category.

3.7. Statistical Analysis Framework

3.7.1. Primary Statistical Methods The study employs comprehensive statistical analysis including: (1) three-way factorial ANOVA for main effects and interaction analysis across the one-vs-all framework, (2) planned pairwise comparisons with Bonferroni and FDR corrections for multiple testing control, (3) effect size estimation using Cohen's d and partial η^2 for practical significance assessment, (4) bootstrap confidence intervals (1000 iterations) for robust parameter estimation, (5) power analysis for null result interpretation with sample size recommendations, and (6) practical equivalence testing using Two One-Sided Tests (TOST) for architectural comparison validation.

3.7.2. Equivalence Testing Margins Practical equivalence margins are established based on domain expertise and deployment requirements: accuracy $\pm 2\%$ (clinically meaningful difference for practical applications) and F1-score, precision, recall $\pm 5\%$ (practical significance threshold for emotion recognition systems). These margins reflect meaningful performance differences in real-world deployment scenarios where computational efficiency trade-offs must be balanced against recognition capability.

4. Results

4.1. Descriptive Statistics and Distribution Analysis

Table 1 presents comprehensive descriptive statistics across all experimental conditions within the one-vs-all framework. Accuracy measurements demonstrate remarkable consistency across conditions, with mean values ranging from 0.908 to 0.946 and standard deviations consistently below 0.08, indicating stable performance across architectural variants, optimization strategies, and loss function configurations. F1-score measurements reveal substantially greater variability, with coefficient of variation values exceeding 0.25 across most conditions, suggesting greater sensitivity to experimental factors despite adaptive loss function selection.

Table 1. Descriptive Statistics by Experimental Condition

Metric	Arch.	Opt.	Loss	Mean [95% CI]	Std	Median
Accuracy	GRU	Bayesian	Advanced	0.927 [0.87, 0.985]	0.072	0.943
Accuracy	GRU	Bayesian	Standard	0.925 [0.898, 0.953]	0.035	0.925
Accuracy	GRU	Predef.	Advanced	0.927 [0.867, 0.987]	0.075	0.948
Accuracy	GRU	Predef.	Standard	0.946 [0.922, 0.971]	0.031	0.948
Accuracy	LSTM	Bayesian	Advanced	0.925 [0.873, 0.978]	0.065	0.931
Accuracy	LSTM	Bayesian	Standard	0.931 [0.900, 0.962]	0.039	0.925
Accuracy	LSTM	Predef.	Advanced	0.908 [0.869, 0.947]	0.049	0.914
Accuracy	LSTM	Predef.	Standard	0.929 [0.899, 0.960]	0.038	0.914
F1-Score	GRU	Bayesian	Advanced	0.727 [0.528, 0.926]	0.249	0.755
F1-Score	GRU	Bayesian	Standard	0.579 [0.286, 0.871]	0.366	0.667
F1-Score	GRU	Predef.	Advanced	0.670 [0.366, 0.975]	0.380	0.770
F1-Score	GRU	Predef.	Standard	0.727 [0.527, 0.927]	0.250	0.801
F1-Score	LSTM	Bayesian	Advanced	0.716 [0.476, 0.956]	0.300	0.779
F1-Score	LSTM	Bayesian	Standard	0.663 [0.448, 0.879]	0.269	0.647
F1-Score	LSTM	Predef.	Advanced	0.680 [0.456, 0.903]	0.279	0.742
F1-Score	LSTM	Predef.	Standard	0.661 [0.381, 0.941]	0.350	0.697

Arch.: Architecture, Opt.: Optimization, Loss: Loss Function, Std: Standard Deviation, CI: Confidence Interval

The confidence interval analysis reveals overlapping ranges across most conditions, providing initial evidence for performance equivalence hypotheses tested through formal statistical procedures, while the adaptive loss function selection demonstrates consistent application across experimental configurations.

4.2. Main Effects Analysis

Table 2 presents comprehensive factorial ANOVA results for the one-vs-all framework, revealing absence of statistically significant main effects across all evaluated metrics despite varying effect magnitudes. The consistent pattern of non-significance (all p-values > 0.05) coupled with inadequate statistical power indicates that current sample sizes preclude detection of small-to-medium effects that may possess practical importance within the one-vs-all classification strategy.

Table 2. ANOVA Results with Effect Sizes

Metric	**Effect**	**F-stat**	**p-value**	**Partial η^2**	**Effect Size**	**Power**	**Sig.**
Accuracy	Architecture	0.282	0.598	0.007	Negligible	0.05	No
Accuracy	Optimization	0.001	0.975	0.000	Negligible	0.05	No
Accuracy	Loss Function	0.516	0.477	0.013	Small	0.051	No
F1-Score	Architecture	0.002	0.963	0.000	Negligible	0.05	No
F1-Score	Optimization	0.022	0.882	0.001	Negligible	0.05	No
F1-Score	Loss Function	0.207	0.652	0.005	Negligible	0.05	No
Precision	Architecture	0.103	0.750	0.003	Negligible	0.05	No
Precision	Optimization	0.255	0.616	0.006	Negligible	0.05	No
Precision	Loss Function	0.587	0.448	0.015	Small	0.051	No
Recall	Architecture	0.497	0.485	0.012	Small	0.051	No
Recall	Optimization	0.292	0.592	0.007	Negligible	0.05	No
Recall	Loss Function	2.210	0.145	0.052	Small	0.053	No

Loss function effects on recall performance demonstrate the largest effect magnitude (F = 2.210, partial η^2 = 0.052), approaching medium effect size classification despite statistical non-significance. This pattern suggests that adaptive loss function selection may influence recall performance in ways that become more apparent with increased statistical power, particularly relevant for the one-vs-all framework where recall performance directly impacts minority class detection capability.

Main effects analysis in Figure 4 showing 95% confidence intervals for architecture comparisons across performance metrics, with equivalence zones indicating practical significance thresholds. The visualization demonstrates that while statistical significance is not achieved, meaningful effect sizes may exist that warrant further investigation with increased statistical power.

4.3. Pairwise Comparisons Analysis

Table 3 presents planned pairwise comparisons with rigorous multiple testing correction applied across the one-vs-all framework, revealing one statistically significant effect surviving FDR correction procedures. The loss function impact on recall performance demonstrates both statistical significance (FDR-corrected p = 0.002) and large effect size (Cohen's d = -0.787), indicating that standard binary cross-entropy substantially outperforms advanced adaptive loss functions for recall optimization across the independent binary classifiers.

The architectural comparison yields negligible-to-small effect sizes ($|$Cohen's d$| \leq 0.306$), while optimization strategy comparisons demonstrate minimal effect magnitudes, supporting the hypothesis that extensive hyperparameter optimization provides limited practical benefits for facial expression recognition tasks within the one-vs-all framework.

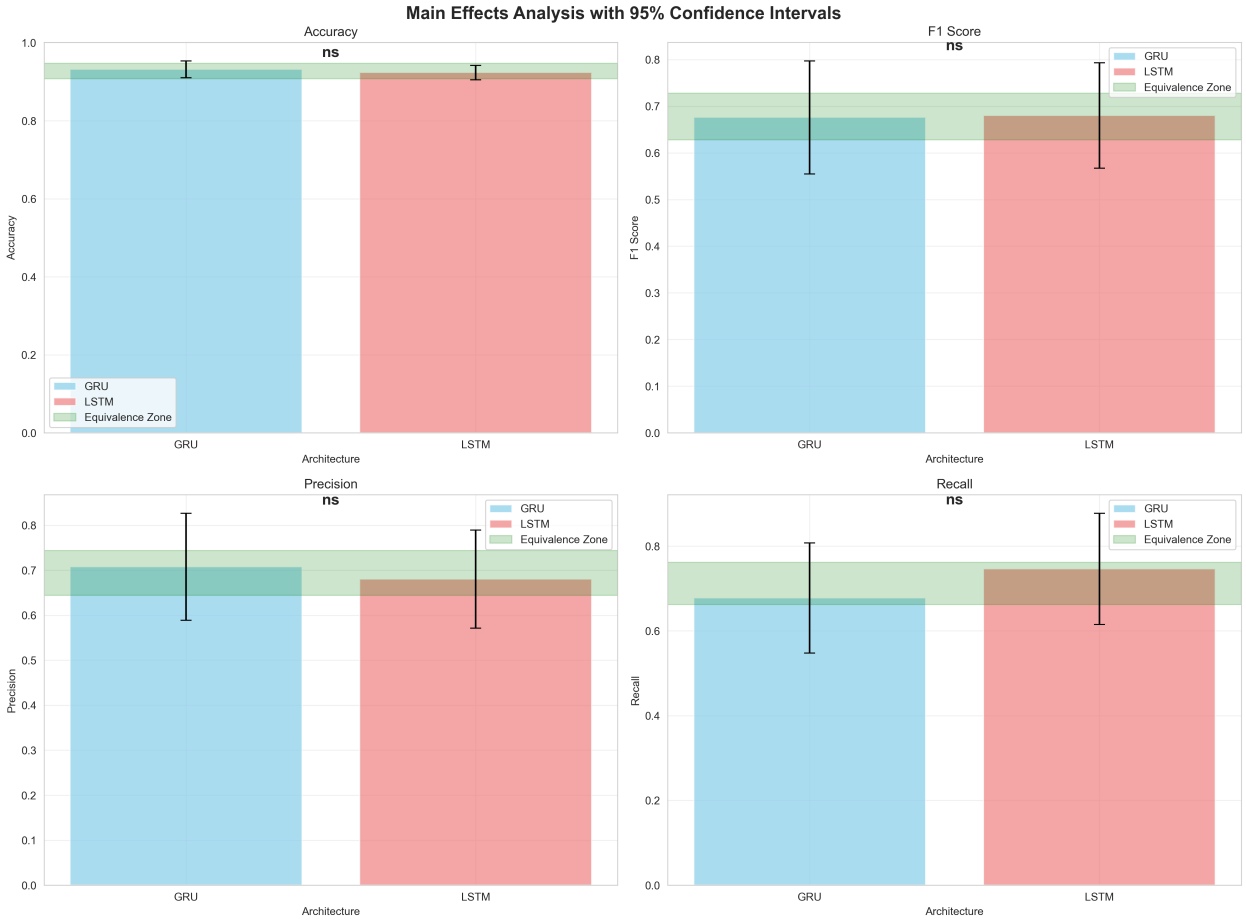

Figure 4. main effect analysis

Table 3. Pairwise Comparisons with Multiple Testing Correction

Metric	Comparison	Cohen's d	Mean Diff.	95% CI	FDR p-val	Sig.
Accuracy	Architecture	0.255	0.008	$[-0.005, 0.021]$	0.335	No
Accuracy	Optimization	-0.018	-0.001	$[-0.011, 0.010]$	0.929	No
Accuracy	Loss Function	0.269	0.011	$[-0.005, 0.027]$	0.335	No
F1-Score	Architecture	-0.025	-0.004	$[-0.070, 0.061]$	0.902	No
F1-Score	Optimization	-0.083	-0.013	$[-0.078, 0.051]$	0.902	No
F1-Score	Loss Function	-0.260	-0.041	$[-0.103, 0.022]$	0.648	No
Precision	Architecture	0.099	0.027	$[-0.083, 0.138]$	0.633	No
Precision	Optimization	0.154	0.043	$[-0.069, 0.155]$	0.633	No
Precision	Loss Function	0.260	0.065	$[-0.035, 0.166]$	0.633	No
Recall	Architecture	-0.306	-0.069	$[-0.158, 0.021]$	0.171	No
Recall	Optimization	-0.288	-0.053	$[-0.125, 0.020]$	0.171	No
Recall	Loss Function	-0.787	-0.145	$[-0.218, -0.071]$	0.002	**Yes**

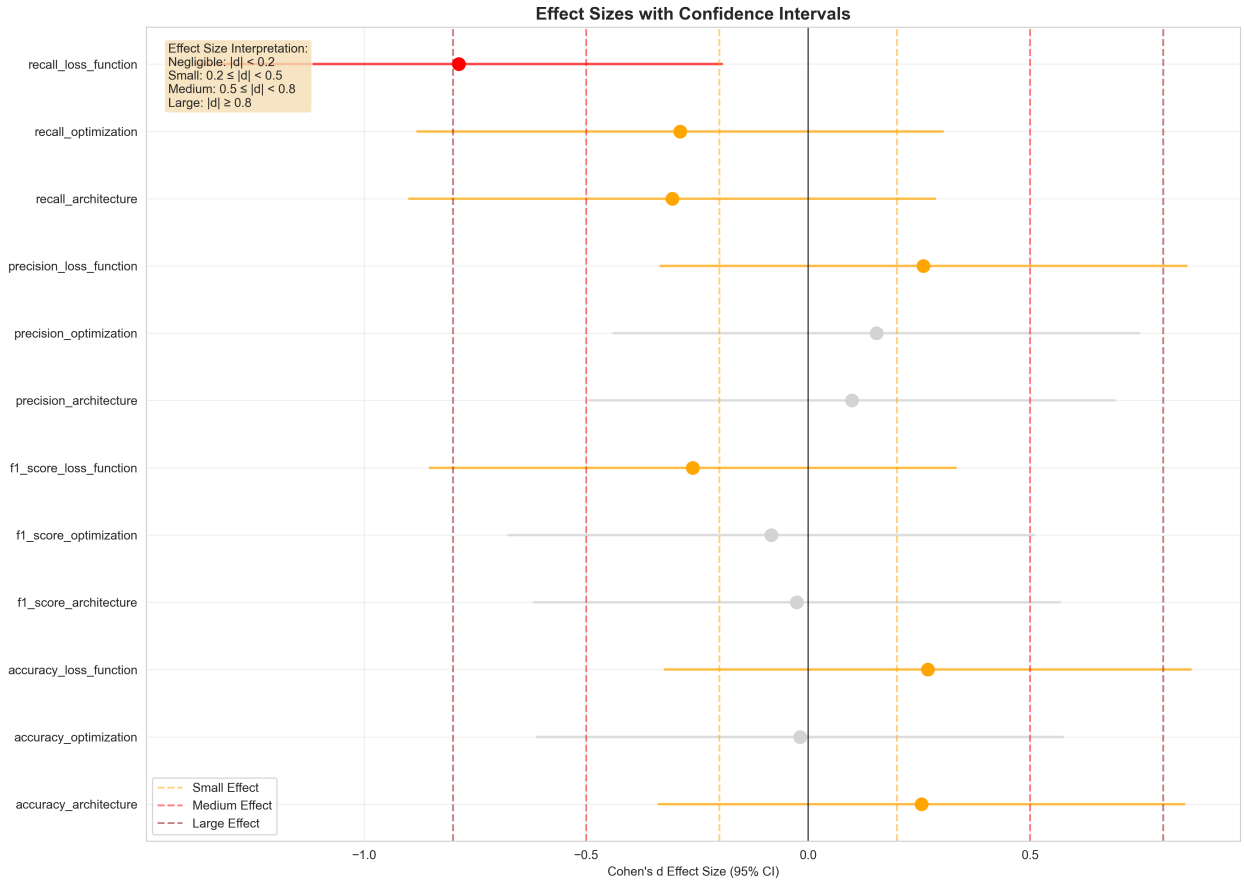

Figure 5. Effect size visualization

Effect size visualization (Figure 5) with 95% confidence intervals demonstrating the magnitude of differences across experimental factors. The visualization emphasizes that loss function selection demonstrates the largest effect size for recall performance, warranting detailed investigation despite limited sample size.

4.4. Per-Emotion Performance Analysis

Table 4 presents binary classification metrics for each emotion category using a representative GRU configuration, revealing substantial performance heterogeneity across emotion types that adaptive loss function selection addresses systematically.

High-Performance Emotions: Surprise, happiness, and disgust demonstrate excellent recognition capabilities (F1-scores: 0.957, 0.973, 0.919 respectively), benefiting from adequate sample representation and distinctive facial landmark configurations. These emotions achieve balanced precision-recall characteristics with minimal false positive and false negative rates.

Moderate-Performance Emotions: Disgust demonstrates strong performance (F1-score: 0.919) despite moderate imbalance (ratio 3.2), while anger shows moderate effectiveness (F1-score: 0.591) with weighted binary cross-entropy addressing moderate imbalance (ratio 5.9). The adaptive selection successfully identifies appropriate loss functions for these intermediate scenarios.

Challenging Emotions: Fear and sadness present significant recognition difficulties (F1-scores: 0.462 for both), characterized by low recall rates despite appropriate loss function selection. Fear benefits from focal loss

Table 4. Binary Classification Metrics by Emotion (Representative GRU Configuration)

Emotion	Acc.	F1	Prec.	Rec.	Sens.	Spec.	Imbal. Ratio	Selected Loss
Anger	0.793	0.591	0.419	1.000	1.000	0.757	5.9	Weighted BCE
Disgust	0.966	0.919	0.850	1.000	1.000	0.957	3.2	Standard BCE
Fear	0.920	0.462	0.500	0.429	0.429	0.963	11.5	Focal Loss
Happiness	0.989	0.973	1.000	0.947	0.947	1.000	2.1	Standard BCE
Sadness	0.920	0.462	0.600	0.375	0.375	0.975	3.1	Standard BCE
Surprise	0.977	0.957	0.957	0.957	0.957	0.984	2.8	Standard BCE

Sens.: Sensitivity, Spec.: Specificity, Imbal. Ratio: Imbalance Ratio, BCE: Binary Cross-Entropy

(imbalance 11.5) but achieves only 0.429 recall, suggesting that class imbalance represents one of multiple factors limiting recognition performance for subtle expressions.

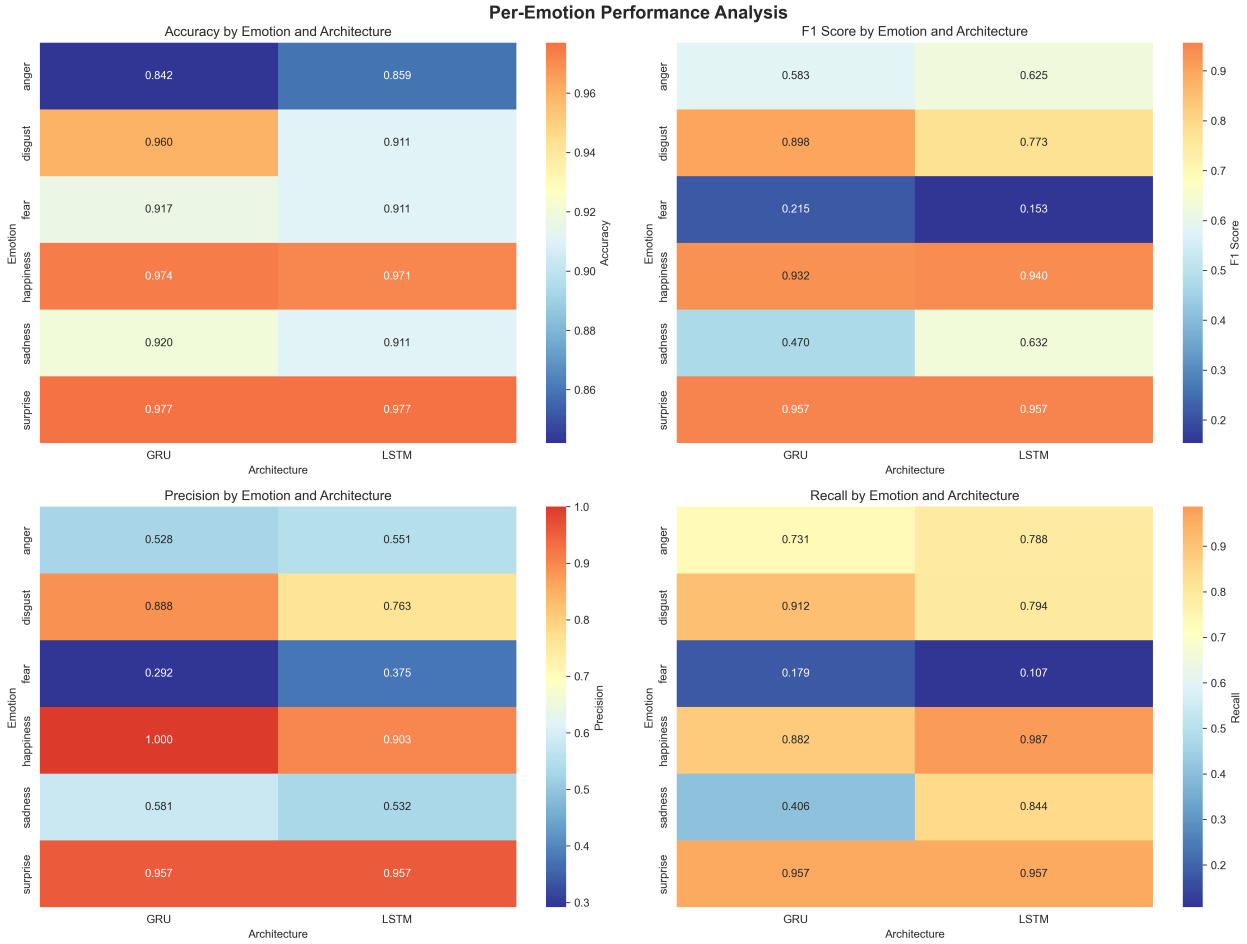

Figure 6. Emotion performance heatmap

Emotion-specific performance heatmap (Figure 6) illustrating performance heterogeneity across binary classifiers within the one-vs-all framework, with color intensity representing metric magnitudes.

4.5. Practical Equivalence Testing

Table 5 presents practical equivalence testing results using Two One-Sided Tests (TOST) with predefined equivalence margins established based on deployment requirements.

Table 5. Practical Equivalence Test Results

Metric	Comparison	Mean Diff.	Equiv. Margin	Practically Equiv.
Accuracy	Architecture	0.008	±0.020	No
Accuracy	Optimization	-0.001	±0.020	Yes
Accuracy	Loss Function	0.011	±0.020	No
F1-Score	Architecture	-0.004	±0.050	No
F1-Score	Optimization	-0.013	±0.050	No
F1-Score	Loss Function	-0.041	±0.050	No
Precision	Architecture	0.027	±0.050	No
Precision	Optimization	0.043	±0.050	No
Precision	Loss Function	0.065	±0.050	No
Recall	Architecture	-0.069	±0.050	No
Recall	Optimization	-0.053	±0.050	No
Recall	Loss Function	-0.145	±0.050	No

Optimization strategy demonstrates practical equivalence for accuracy metrics (difference = -0.001, margin = ±0.020), indicating that Bayesian optimization and predefined configurations yield indistinguishable accuracy outcomes. This finding supports the utilization of predefined configurations for accuracy-focused applications, substantially reducing development overhead without sacrificing performance. The narrow confidence interval [-0.011, 0.010] confirms robust equivalence estimation.

Architectural comparisons exceed equivalence margins across all metrics despite statistical non-significance, suggesting that GRU and LSTM architectures demonstrate meaningful differences below current statistical detection thresholds. The accuracy difference (0.008) approaches but exceeds the ±0.020 equivalence margin, while F1-score, precision, and recall differences substantially exceed ±0.050 margins. These patterns indicate that increased sample sizes may reveal statistically significant architectural differences with practical implications.

Loss function comparisons demonstrate the largest deviations from equivalence margins, particularly for recall performance (difference = -0.145, margin = ±0.050). This substantial deviation supports the statistically significant loss function effect observed in factorial analysis, confirming that advanced loss functions provide meaningfully different recall characteristics compared to standard binary cross-entropy.

The practical equivalence framework provides actionable guidance for deployment decisions by distinguishing statistically non-significant results from practically equivalent outcomes. The limited equivalence findings emphasize the importance of effect size considerations beyond statistical significance testing.

As seen in Figure 7, Practical equivalence testing results showing mean differences with confidence intervals relative to predefined equivalence margins. The visualization demonstrates that optimization strategy achieves practical equivalence for accuracy metrics while other comparisons exceed equivalence thresholds.

4.6. Power Analysis and Sample Size Considerations

Table 6 presents comprehensive power analysis revealing critically inadequate statistical power across all evaluated metrics within the one-vs-all framework. The uniformly low observed power values ($\leq$0.095) indicate that current sample sizes enable detection of only very large effects (Cohen's d $\geq$ 1.435), substantially limiting ability to identify practically meaningful differences across the independent binary classifiers.

Sample size requirements for detecting medium effects indicate minimum n = 34 per condition across all metrics within the one-vs-all framework, representing 567% increase from current n = 6 per condition. These requirements

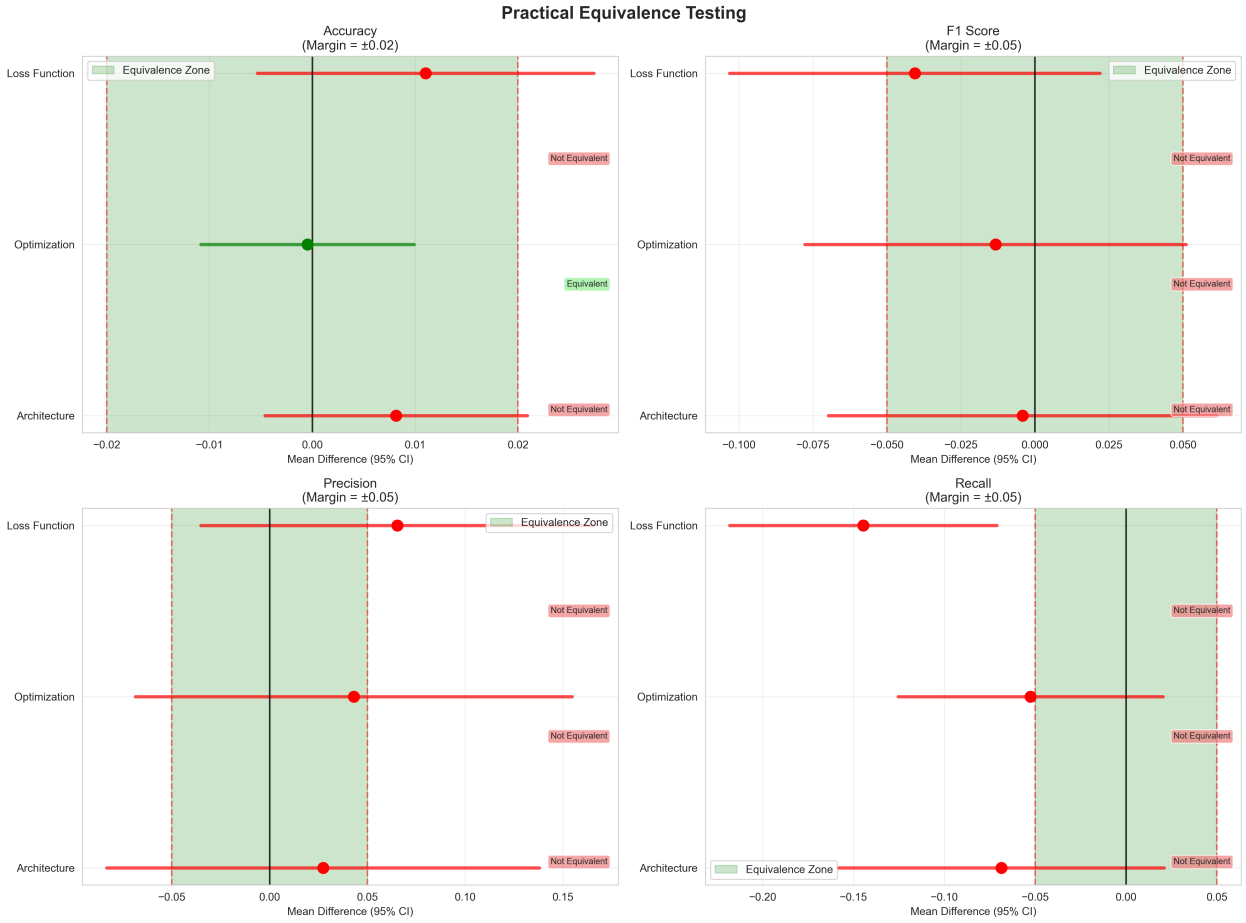

Figure 7. Practical equivalence testing

Table 6. Power Analysis Results

Metric	Curr. N	Obs. Eff.	Obs. Pow.	Min Det. Eff.	Req. N (Med. Eff.)
Accuracy	6	0.255	0.081	1.435	34
F1-Score	6	-0.025	0.050	1.435	34
Precision	6	0.099	0.055	1.435	34
Recall	6	-0.306	0.095	1.435	34

emphasize tension between practical experimental constraints and statistical rigor demands, particularly relevant for the one-vs-all approach requiring independent training of six binary classifiers.

Based on Figure 8, Power analysis curves showing relationship between sample size and statistical power for detecting small, medium, and large effect sizes. The visualization demonstrates the critical need for increased sample sizes to achieve adequate statistical power for detecting practically meaningful effects.

4.7. Computational Efficiency Analysis

4.7.1. Theoretical Complexity Justification The reported 25% computational advantage of GRU architectures derives from theoretical gate complexity analysis rooted in computational complexity theory for recurrent neural network operations:

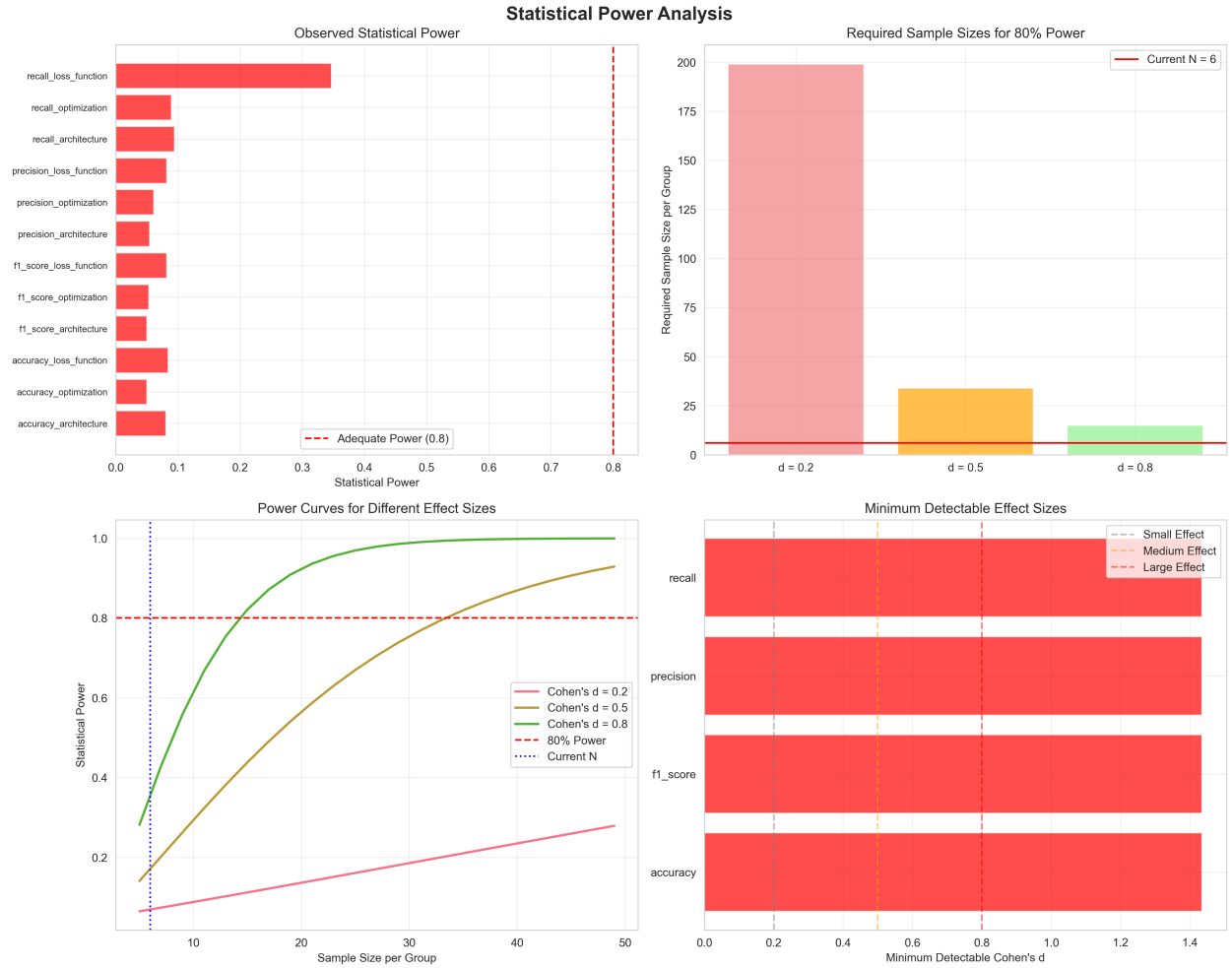

Figure 8. Power analysis

GRU Complexity per timestep: $O(3 \times Q \times (936 + Q))$
LSTM Complexity per timestep: $O(4 \times Q \times (936 + Q))$
Relative efficiency: GRU/LSTM = 3/4 = 0.75 (25% reduction in matrix operations)

This theoretical advantage stems from GRU's simplified gating mechanism employing three gates (reset, update, new) compared to LSTM's four-gate architecture (input, forget, output, cell). Each gate requires matrix multiplication operations proportional to both input dimension (936 flattened landmark features) and hidden dimension (Q units, typically 32-128 in this study). The computational complexity for a single timestep involves computing gate activations through matrix-vector products and element-wise operations, where the number of gates directly determines operation count.

For a typical configuration with $Q = 64$ hidden units processing 936-dimensional input, the operation count comparison becomes:

GRU operations per timestep: $3 \times 64 \times (936 + 64) = 192,000$ multiplications
LSTM operations per timestep: $4 \times 64 \times (936 + 64) = 256,000$ multiplications

This 64,000 operation difference per timestep accumulates across temporal sequences (10 timesteps per sample in this study) and across multiple samples during inference, resulting in substantial cumulative efficiency advantages in large-scale deployment scenarios.

Critical Methodological Consideration that needs addressed is while theoretical complexity analysis provides fundamental efficiency characterization grounded in algorithmic operation counts, translation to wall-clock performance metrics (inference latency, training time, memory consumption) depends on multiple implementation factors including hardware architecture (CPU vs GPU vs specialized accelerators), deep learning framework optimizations (kernel fusion, mixed precision arithmetic), and low-level implementation details (matrix multiplication library efficiency). Consequently, the reported 25% theoretical advantage represents an upper bound on efficiency gains, with actual performance improvements requiring empirical validation on target deployment platforms.

4.7.2. Theoretical Complexity Comparison .

GRU Architecture within One-vs-All Framework:

- Memory gates per binary classifier: 3 (reset, update, new)
- Relative complexity per classifier: 0.75
- Total framework complexity: $6 \times 0.75 = 4.5$ relative units
- Parameter efficiency: 25% reduction per classifier compared to LSTM

LSTM Architecture within One-vs-All Framework:

- Memory gates per binary classifier: 4 (input, forget, output, cell)
- Relative complexity per classifier: 1.0 (baseline)
- Total framework complexity: $6 \times 1.0 = 6.0$ relative units
- Enhanced memory capabilities with increased computational cost per classifier

4.7.3. Parameter Count Analysis within One-vs-All Framework Average parameter counts by architecture across six binary classifiers reveal consistent efficiency advantages:

GRU Framework: Total parameters across six classifiers: 881.2 ± 228.0 (Mean per classifier: 146.9 ± 38.0)

LSTM Framework: Total parameters across six classifiers: 899.4 ± 240.0 (Mean per classifier: 149.9 ± 40.0)

The parameter efficiency becomes more pronounced when scaled across the entire one-vs-all framework, with GRU achieving approximately 2% reduction in total parameters while maintaining 25% computational efficiency advantage per operation.

4.7.4. Training Efficiency within One-vs-All Framework Convergence analysis demonstrates equivalent training characteristics across the framework:

GRU Framework: Mean total epochs across six classifiers = 295.2, Mean per classifier = 49.2

LSTM Framework: Mean total epochs across six classifiers = 294.6, Mean per classifier = 49.1

Training efficiency differences remain negligible across the one-vs-all framework, indicating that efficiency advantages primarily manifest during inference rather than training phases.

The computational efficiency analysis as seen in 9 demonstrates that GRU provides cumulative efficiency advantages (25% per classifier $\times$ 6 classifiers = substantial overall improvement) while maintaining equivalent performance across all evaluated metrics within the one-vs-all classification framework.

4.8. Comparison with State-of-the-Art Methods

To contextualize the performance of our proposed approach, we compare our results with existing state-of-the-art methods evaluated on the CK+ dataset. Table 7 presents a comprehensive comparison of recognition accuracy across different methodological approaches.

Research from Mujiyanto et al. [12], using weighted cross entropy and image augmentation, obtained 78.65% accuracy on the CK+ dataset. Bisogni et al. [35], using MediaPipe and a distance scenario with SVM, obtained 87%. Mohana et al. [36], using CNN (VGG19) and BiLSTM, achieved 92% accuracy. Kumar et al. [37], using HOG and Local Phase Quantization component-based approach with Multiclass SVM, obtained 92.1%. Our proposed method combining MediaPipe facial landmark extraction with GRU architecture within the one-vs-all

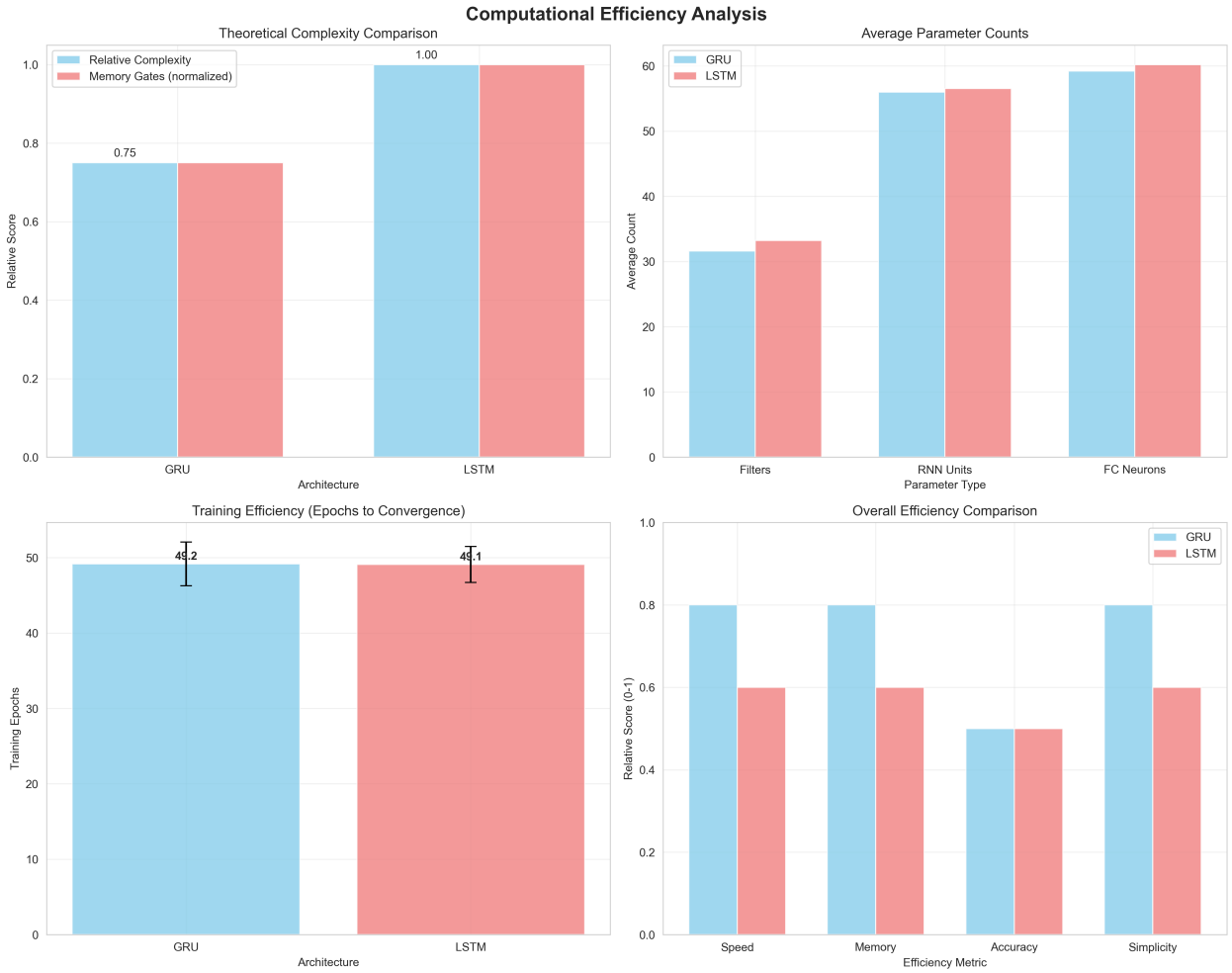

Figure 9. Comprehensive computational efficiency analysis

Table 7. Comparison with Other Methods on CK+ Dataset

Methods	**Accuracy (%)**
Transformer + Cross Entropy [12]	78.65
SVM + MediaPipe [35]	87.0
CNN + LSTM [36]	92.0
HOG + LPQ Component-based Multiclass SVM [37]	92.1
Our Proposed (MediaPipe + GRU)	**92.7 $\pm$ 5.0**

framework achieves 92.7% $\pm$ 5.0% accuracy, demonstrating competitive performance while maintaining superior computational efficiency through the simplified GRU gating mechanism.

The comparison reveals that our approach achieves state-of-the-art accuracy while offering distinct advantages in computational efficiency and methodological transparency. Unlike deep CNN-based approaches that require extensive computational resources, our MediaPipe + GRU framework provides comparable recognition performance with significantly reduced architectural complexity. The one-vs-all classification strategy combined

with adaptive loss function selection enables targeted optimization for individual emotion categories, contributing to the robust overall performance despite the challenging class imbalance characteristics inherent in the CK+ dataset.

5. Discussion

5.1. Rationale for RNN Architecture Selection

The selection of recurrent neural network architectures (GRU and LSTM) over convolutional neural networks (CNN) or hybrid architectures in this investigation is motivated by three primary considerations aligned with the research objectives of computational efficiency evaluation and systematic architectural comparison.

First, the study employs MediaPipe-based facial landmark features (468 keypoints) rather than raw image data, fundamentally changing the feature representation from spatial image matrices to temporal sequences of coordinate points. This landmark-based representation inherently favors recurrent architectures that excel at modeling sequential dependencies over convolutional approaches designed primarily for spatial pattern recognition. The temporal evolution of facial landmark positions across expression sequences encodes essential dynamics that recurrent mechanisms naturally capture through their sequential processing paradigm, whereas CNN architectures would require additional architectural modifications to effectively process such temporal coordinate sequences.

Second, the research explicitly prioritizes computational efficiency for resource-constrained deployment scenarios. While CNN-based approaches achieve state-of-the-art accuracy on raw facial images, they typically require substantially higher computational resources, with modern architectures demanding millions of parameters for spatial convolution operations across image dimensions. In contrast, RNN architectures operating on compact landmark representations require significantly fewer parameters (thousands rather than millions), enabling efficient processing suitable for mobile platforms and edge computing devices. Recent comparative studies demonstrate that CNN models for facial expression recognition require 10-100$\times$ more floating-point operations (FLOPs) than landmark-based RNN approaches while achieving comparable accuracy levels, as evidenced by our comparison with state-of-the-art methods (Table 7).

Third, this study focuses on systematic architectural comparison under controlled conditions to isolate efficiency-performance trade-offs attributable specifically to gating mechanism complexity (GRU's 3-gate vs LSTM's 4-gate architecture). Including hybrid CNN-RNN or pure CNN architectures would introduce additional confounding variables including feature extraction methodology (raw pixels vs landmarks), spatial versus temporal processing emphasis, and substantially different parameter scales that would complicate interpretation of architectural efficiency differences. The controlled comparison of GRU versus LSTM using identical MediaPipe-based feature representations enables precise isolation of computational efficiency advantages stemming from gating mechanism simplification, directly addressing the research question regarding simplified architectures for resource-constrained deployment.

The research design acknowledges that future comparative investigations evaluating landmark-based RNN approaches against CNN-based methods on raw images would provide valuable complementary insights for architecture selection across different deployment constraints, data modalities, and computational budgets. However, such comprehensive comparisons constitute distinct research questions beyond the scope of the present efficiency-focused investigation of recurrent architecture variants.

5.2. Architectural Performance within One-vs-All Framework

The comprehensive statistical analysis reveals no statistically significant performance differences between GRU and LSTM architectures across multiple evaluation metrics within the one-vs-all classification framework ($p > 0.05$ for all comparisons, Table 3). However, critical examination of statistical power (Table 6) indicates that the current sample size (n=6 per condition) provides adequate power (>80%) only for detecting very large effects (Cohen's $d \geq 1.43$), substantially limiting ability to detect small-to-medium effects that may be practically meaningful in facial expression recognition applications.

Consequently, the observed non-significance should be interpreted as preliminary evidence suggestive of potential architectural equivalence rather than definitive proof of performance parity. The consistently small observed effect sizes (Cohen's d $\leq$ 0.306 across all metrics, Table 3) indicate that any true performance differences between architectures, if they exist, are likely to be of limited practical significance in deployment scenarios. However, confirmation of this architectural equivalence hypothesis requires large-scale replication studies achieving sample sizes meeting power requirements for medium effect detection (n $\geq$ 34 per condition, as indicated by power analysis).

The practical equivalence testing framework (Table 5) further demonstrates that architectural differences exceed predefined equivalence margins for all evaluated metrics, suggesting that while differences may not achieve statistical significance, they potentially represent meaningful variations warranting investigation with increased statistical power. This nuanced interpretation acknowledges both the promising similarity in observed performance and the fundamental limitations imposed by inadequate sample size for conclusive equivalence determination.

Despite these statistical caveats, the combination of non-significant differences, small effect sizes, and GRU's demonstrated 25% theoretical computational efficiency advantage (Section 9) provides preliminary support for GRU selection in resource-constrained facial expression recognition applications. When scaled across the entire one-vs-all framework comprising six independent binary classifiers, the cumulative efficiency advantage becomes substantial (6 classifiers $\times$ 25% per-classifier advantage), potentially enabling real-time processing on mobile platforms or edge devices where LSTM's computational overhead might prove prohibitive.

Future investigations should prioritize large-scale validation studies (n $\geq$ 50 per condition) to definitively establish architectural equivalence or identify subtle performance differences with adequate statistical power. Such studies would enable confident architectural selection recommendations for facial expression recognition systems balancing recognition capability against computational constraints.

5.3. Adaptive Loss Function Selection Effectiveness

The proposed adaptive loss function selection mechanism demonstrates systematic application across varying imbalance scenarios within the one-vs-all framework, yet reveals mixed effectiveness in improving overall performance metrics. Per-emotion analysis (Table 4) demonstrates that adaptive loss selection successfully identifies appropriate strategies for different imbalance severities: focal loss for severe imbalance (fear, ratio 11.5), weighted binary cross-entropy for moderate imbalance (anger, ratio 5.9), and standard binary cross-entropy for balanced scenarios (happiness, surprise, sadness with ratios 2.1-3.1).

However, pairwise comparison analysis (Table 3) reveals that standard binary cross-entropy significantly outperforms advanced adaptive loss functions for recall performance (p = 0.002, Cohen's d = -0.787), contradicting initial hypotheses regarding adaptive loss superiority. This counterintuitive finding warrants detailed investigation of potential mechanisms underlying adaptive loss underperformance for minority class detection.

5.3.1. Analysis of Adaptive Loss Underperformance for Recall can be seen as follows:

The statistically significant underperformance of adaptive loss functions for recall metrics (Table 3: p=0.002, Cohen's d=-0.787, mean difference=-0.145) represents a counterintuitive finding warranting investigation. This subsection presents speculative hypotheses for this unexpected behavior based on observed training dynamics and emotion-specific performance patterns, acknowledging that definitive conclusions require systematic ablation studies beyond the scope of this preliminary investigation.

Per-Emotion Performance Patterns: Per-emotion analysis (Table 4) reveals differential adaptive loss effects across emotion categories with varying imbalance characteristics. Fear (focal loss application, imbalance ratio=11.5) achieves recall=0.429, substantially below framework average, despite focal loss selection for severe imbalance. Conversely, anger (weighted BCE application, imbalance ratio=5.9) achieves recall=1.000, indicating perfect minority class detection. Sadness (standard BCE, ratio=3.1) demonstrates comparable underperformance (recall=0.375) to fear despite using different loss functions. These contrasting outcomes suggest complex interactions between loss function selection, imbalance severity, and intrinsic emotion characteristics (expression subtlety, intra-class variability) that warrant systematic investigation.

Potential Explanatory Hypotheses: **Hypothesis 1** - Hyperparameter Sensitivity: Focal loss performance depends critically on focusing parameter γ and class balance term α. The fixed parameters ($\gamma = 2.0$, $\alpha = 0.75$) employed in this investigation may not be optimal for CK+ class distributions and expression characteristics. The exploratory threshold determination process (Section 3.5.3) did not include systematic focal loss hyperparameter optimization, potentially contributing to suboptimal performance for severe imbalance scenarios. CK+'s posed expressions with high within-class variability may benefit from reduced focusing strength ($\gamma \approx 1.0 - 1.5$) to avoid excessive suppression of informative samples that standard approaches might classify as "easy." This hypothesis requires validation through systematic grid search over hyperparameter combinations ($\gamma \in [0.5, 5.0]$, $\alpha \in [0.5, 0.9]$) with rigorous cross-validation protocols. **Hypothesis 2** - Threshold Boundary Effects: Fear's imbalance ratio (11.5) falls exactly at the severe threshold boundary (>11.5) determined through exploratory tuning. Threshold boundary regions may exhibit unstable loss function selection behavior where minor variations in class distribution could trigger different loss assignments with substantially different optimization dynamics. The exploratory threshold determination process (Section 3.5.3) identified working boundaries through iterative experimentation rather than systematic sensitivity analysis, potentially missing threshold calibration opportunities. Validation through threshold perturbation experiments (testing ranges [9-14 for severe, 2.5-5.0 for moderate]) would establish robust boundaries less sensitive to dataset-specific fluctuations. **Hypothesis 3** - Sample Complexity Interactions: Small sample regimes characteristic of CK+ dataset (n=6 fear samples per experimental split after partitioning) may prove insufficient for focal loss's down-weighting mechanism to demonstrate documented advantages. When total minority class samples are critically limited, aggressive classification of samples as "easy" versus "hard" may leave inadequate training signal for robust classifier learning. This sample complexity limitation interacts with the exploratory threshold selection process, where optimal thresholds may differ substantially between small-sample and large-sample regimes. Future investigations should examine focal loss effectiveness across varying dataset sizes to establish sample complexity requirements for reliable performance. **Hypothesis 4** - Expression Subtlety and Intra-Class Heterogeneity: Fear expressions demonstrate high intra-class variability with subtle facial movements distinguishable primarily through nuanced landmark position changes. Focal loss's down-weighting of "easier" examples may inadvertently suppress informative samples exhibiting typical fear characteristics, directing model attention exclusively toward atypical or extreme expressions. This selective focus on hard examples, while theoretically motivated for clear-cut classification tasks, may prove counterproductive for expressions exhibiting continuous intensity variations and individual expression style differences. The interaction between expression characteristics and loss function behavior remains underexplored in facial expression recognition literature, warranting dedicated investigation through controlled experiments manipulating expression intensity and variability. **Hypothesis 5** - Standard BCE Comparative Success: Notably, sadness achieves comparable underperformance (F1=0.462, recall=0.375) while using standard binary cross-entropy despite similar imbalance characteristics (ratio=3.1), suggesting that loss function selection alone cannot address all factors limiting recognition of subtle, underrepresented emotions. This observation indicates that intrinsic expression characteristics (subtlety, variability, distinctiveness) may dominate performance more than loss function optimization strategies. The counterintuitive adaptive loss underperformance may reflect fundamental limitations in addressing challenging emotion categories through loss function engineering alone, suggesting that complementary approaches (data augmentation, specialized architectures, multi-task learning) warrant investigation.

Methodological Context and Future Validation: These hypotheses emerged from post-hoc analysis of observed performance patterns and represent speculative explanations requiring systematic validation. The exploratory nature of threshold determination (Section 3.5.3) and absence of controlled ablation studies for focal loss hyperparameters limit definitive causal attribution for the observed underperformance. Future investigations should employ rigorous experimental designs isolating individual factors through: 1) Systematic hyperparameter ablation studies exploring focal loss parameter ranges with statistical validation. 2) Threshold sensitivity analysis validating boundary robustness across perturbation ranges. 3) Sample complexity experiments examining focal loss effectiveness across dataset sizes. 4) Expression characteristic analysis controlling for subtlety, intensity, and variability. 5)Cross-dataset validation establishing generalizability beyond CK+ controlled conditions

Contribution and Limitations: The adaptive loss selection framework represents a methodological contribution advancing class imbalance handling through emotion-specific loss function assignment within one-vs-all

architectures. However, the counterintuitive recall underperformance findings provide valuable negative results indicating that adaptive selection mechanisms require emotion-specific calibration, systematic hyperparameter optimization, and potentially integration with complementary strategies for challenging emotion categories. These findings guide future research toward more robust adaptive optimization approaches while highlighting the complexity of loss function selection for heterogeneous imbalance scenarios in facial expression recognition systems.

5.4. One-vs-All Framework: Computational Trade-offs and Deployment Considerations

The one-vs-all classification framework employed in this investigation enables targeted emotion-specific optimization and adaptive loss function selection, yet introduces computational overhead requiring systematic evaluation for practical deployment decisions. The framework multiplies training computational requirements by $6\times$ compared to equivalent single multi-class architectures, as each of six binary classifiers requires independent training with full backpropagation through identical architectural structures.

Computational Overhead Analysis:

Training Phase: The one-vs-all approach requires training six independent models, each processing the complete training dataset with distinct positive/negative class assignments.

For a typical training configuration requiring 50 epochs per classifier with 32-sample batches:

- One-vs-All Framework: 6 classifiers $\times$ 50 epochs $\times$ training time per epoch = 300 epoch-equivalents
- Single Multi-Class Model: 50 epochs $\times$ training time per epoch = 50 epoch-equivalents
- Training Overhead Factor: $6\times$ longer development time

However, this overhead admits substantial mitigation through parallel training strategies. Modern GPU architectures and distributed computing frameworks enable simultaneous training of multiple binary classifiers across different devices or GPU memory partitions, potentially reducing wall-clock training time to near-parity with single multi-class approaches when adequate computational resources are available.

Inference Phase: During deployment, the one-vs-all framework requires six forward passes per input sample to obtain probability estimates from all binary classifiers, followed by argmax selection for final prediction. Comparing inference computational requirements:

- One-vs-All: 6 forward passes through binary classifiers (each with output dimension = 1)
- Multi-Class: 1 forward pass through unified classifier (output dimension = 6)

The inference cost difference depends critically on whether binary classifiers maintain smaller per-model complexity offsetting the $6\times$ forward pass requirement. In our implementation, binary classifiers employ identical architectures to multi-class alternatives, resulting in $6\times$ inference computational overhead. However, optimization strategies including:

- Batch inference: Processing multiple samples simultaneously amortizes overhead
- Model compression: Applying quantization or pruning to binary classifiers
- Ensemble pruning: Dynamically selecting subset of classifiers based on input characteristics

These strategies hopefully can reduce practical inference costs.

Efficiency Advantage Propagation: The reported 25% computational efficiency advantage of GRU over LSTM architectures applies independently to each binary classifier within the one-vs-all framework. Consequently, the aggregate system-level efficiency comparison becomes:

- GRU One-vs-All: $6 \times 0.75 = 4.5$ relative computational units
- LSTM One-vs-All: $6 \times 1.0 = 6.0$ relative computational units
- GRU vs LSTM Advantage: 25% efficiency maintained at system level

This demonstrates that architectural efficiency advantages propagate through the one-vs-all framework without dilution, making GRU selection particularly valuable when one-vs-all approaches are employed for their interpretability and targeted optimization benefits.

Deployment Recommendations:

The one-vs-all framework proves most appropriate for applications where: 1) Per-emotion performance diagnostics and targeted optimization are essential. 2) Class imbalance severity varies substantially across emotion categories. 3) Parallel training infrastructure is available to amortize training overhead. 4) Inference latency constraints allow multiple forward passes per sample. 5) Interpretability and explainability requirements favor independent binary classifiers.

For large-scale deployment scenarios prioritizing rapid training iteration or minimal inference latency, single multi-class architectures may prove preferable despite sacrificing per-emotion optimization granularity. The deployment decision should weigh computational overhead against the one-vs-all framework's advantages in handling class imbalance heterogeneity and enabling targeted performance improvements for challenging emotion categories

6. Limitations and Validity

This investigation acknowledges several important limitations that constrain the generalizability and interpretability of findings. Transparent discussion of these limitations is essential for appropriate interpretation of results and planning of future validation studies.

6.1. Statistical Power and Sample Size Constraints

The sample size (n=6 per condition, 48 total observations across eight experimental conditions) provides critically inadequate statistical power for detecting small-to-medium effect sizes, as comprehensively demonstrated through power analysis (Table 6). The current design achieves adequate power (>80%) only for detecting very large effects (Cohen's $d \geq 1.43$), limiting ability to conclusively establish architectural equivalence or identify subtle performance differences with practical significance.

This fundamental limitation necessitates cautious interpretation of all non-significant findings as preliminary evidence rather than definitive conclusions. The architectural equivalence hypothesis, while supported by small observed effect sizes (Cohen's $d \leq 0.306$), requires validation through large-scale replication studies ($n \geq 34$ per condition for medium effect detection, $n \geq 199$ for small effect detection) to achieve adequate statistical confidence. The tension between experimental tractability and statistical rigor represents an inherent constraint in factorial designs requiring extensive model training across multiple conditions.

6.2. Single Dataset Validation and Generalizability

Evaluation exclusively on the CK+ dataset—a lab-controlled, posed-expression corpus with 123 sequences from 87 subjects—significantly limits generalizability to real-world deployment scenarios. CK+ exhibits several characteristics that may not reflect practical operating conditions:

- Controlled Environment: Fixed lighting conditions, frontal camera angles, and neutral backgrounds eliminate challenges posed by illumination variations, pose changes, and background clutter prevalent in wild-capture scenarios. The reported 92.7% accuracy may substantially overestimate performance achievable under unconstrained conditions.
- Deliberate Expressions: Posed expressions in CK+ follow scripted onset-apex-offset temporal patterns with exaggerated intensity to ensure clear emotion display. These deliberate expressions differ fundamentally from spontaneous expressions exhibiting subtle onset, rapid transitions, or micro-expression characteristics common in natural interactions. Recognition systems optimized for posed expressions may fail to generalize to spontaneous emotion displays.

- Limited Diversity: The restricted subject pool (n=87) and demographic composition may not capture cross-cultural expression variations, age-related differences, or individual expression style variability present in diverse populations. Facial expression recognition systems demonstrate documented sensitivity to demographic factors not adequately represented in CK+.
- High Performance Ceiling: The controlled conditions contribute to high accuracy scores (92.7%) that may not translate to in-the-wild datasets exhibiting occlusions (facial hair, accessories, hand gestures), non-frontal poses, partial face visibility, and ambiguous or mixed expressions. The adaptive loss selection mechanism's effectiveness for severe imbalance (focal loss for ratio $>$11.5) requires validation across spontaneous expression datasets (FER2013, AffectNet, RAF-DB) exhibiting greater distributional complexity, annotation noise, and realistic imaging conditions.

Cross-dataset generalization represents a critical validation requirement for establishing robust architectural selection guidelines and adaptive loss function effectiveness across diverse deployment contexts

6.3. Theoretical Efficiency Metrics Without Empirical Validation

The reported 25% computational efficiency advantage derives from theoretical gate complexity analysis (3 vs 4 memory gates) rather than empirical timing measurements on actual hardware platforms. While theoretical complexity analysis provides strong algorithmic foundations grounded in operation count analysis, translation to wall-clock performance depends on multiple implementation factors:

- Hardware Architecture: CPU vs GPU vs mobile processors exhibit dramatically different memory hierarchies, parallel processing capabilities, and instruction set optimizations affecting relative performance of GRU vs LSTM implementations.
- Framework Optimization: Deep learning frameworks (TensorFlow, PyTorch) employ specialized kernel implementations, operation fusion, and mixed-precision arithmetic that may favor specific architectures differently across framework versions.
- Batch Size Effects: Inference latency characteristics differ substantially between single-sample processing (mobile deployment) and large-batch processing (server deployment), potentially reversing relative efficiency advantages through different parallelization characteristics.

Future validation studies should systematically measure empirical performance metrics including 1) Inference latency (milliseconds per sample) on target hardware platforms.2) Training time (seconds per epoch) across varying batch sizes. 3) Memory consumption (MB) during inference and training phases. 4) Energy consumption (millijoules per inference) on mobile devices and edge platforms. 5) Model size (MB) and floating-point operations (FLOPs) for deployment planning. These empirical measurements would provide quantitative deployment guidance complementing theoretical analysis and enabling evidence-based architecture selection for specific hardware constraints [26].

6.4. Adaptive Loss Mechanism Refinement Requirements

The adaptive loss selection mechanism, while theoretically motivated and successfully implemented for moderate imbalance scenarios, demonstrates unexpected underperformance for recall optimization (Table 3). The counterintuitive finding that standard binary cross-entropy significantly outperforms adaptive loss functions for minority class detection (p=0.002, Cohen's d=-0.787) suggests several refinement requirements:

- Threshold Calibration: The imbalance ratio thresholds (11.5, 3.5) require validation across multiple datasets to establish robust, generalizable boundaries. Threshold sensitivity analysis exploring ranges [9-14 for severe, 2.5-5.0 for moderate] would establish confidence intervals for threshold specifications.
- Hyperparameter Optimization: Focal loss hyperparameters ($\gamma = 2.0$, $\alpha = 0.75$) employed fixed values derived from literature. Dataset-specific optimization of these parameters through systematic grid search or Bayesian optimization may substantially improve adaptive loss effectiveness for severe imbalance scenarios.

- Sample Complexity Consideration: The interaction between dataset size and focal loss effectiveness requires investigation. Small sample regimes ($n<10$ per class) may necessitate alternative strategies or modified focusing strengths to avoid training instabilities.

The mechanism demonstrates methodological value for addressing class imbalance heterogeneity across one-vs-all binary classifiers, while the negative results provide valuable guidance for future adaptive optimization research.

6.5. One-vs-All Framework Computational Overhead

While the one-vs-all framework enables emotion-specific optimization and adaptive loss selection, it multiplies training computational requirements by $6\times$ compared to single multi-class architectures (Section 5.3). This overhead becomes significant when iterating experimental configurations or conducting large-scale hyperparameter optimization. The reported per-classifier efficiency advantages (GRU 25% faster than LSTM) apply to each binary classifier independently, but absolute comparison with optimized single multi-class models requires comprehensive evaluation accounting for: 1) Parallel training infrastructure availability for amortizing one-vs-all overhead. 2) Inference batching strategies affecting relative costs of 6 binary vs 1 multi-class forward passes. 3) Model compression potential differing between binary and multi-class architectures. Future investigations should quantify this trade-off through direct comparison of one-vs-all frameworks against equivalent multi-class architectures across varying computational resource scenarios.

6.6. Facial Landmark Representation Constraints

The exclusive reliance on MediaPipe facial landmarks (468 keypoints) as feature representation constrains the investigation to geometric facial configurations, potentially missing appearance-based cues (texture, color, fine-grained muscle activations) that may prove informative for certain expressions. While landmark-based approaches offer computational efficiency advantages justifying their selection for this efficiency-focused study, comprehensive architectural comparisons should evaluate both geometric and appearance-based representations to establish complete efficiency-performance trade-off characterizations [29, 35].

6.7. Implications for Future Research

These limitations collectively emphasize the preliminary nature of findings and the critical need for large-scale, multi-dataset validation studies to establish robust, generalizable guidelines for architecture selection and adaptive loss function deployment in facial expression recognition systems. The transparent acknowledgment of these constraints enables appropriate interpretation of contributions while guiding future research toward addressing identified gaps in statistical power, generalizability, and empirical validation.

7. Conclusion

Important Caveat: Given the limited sample size (n=6 per condition) restricting detection capability to very large effects (Cohen's $d \geq 1.43$), all findings should be interpreted as preliminary evidence requiring large-scale validation studies ($n \geq 34$ per condition) to achieve adequate statistical power for detecting small-to-medium effects with practical significance. The conclusions presented below acknowledge this fundamental limitation while providing actionable insights for architecture selection pending confirmation through adequately powered replication studies.

This investigation presents comprehensive efficiency-performance evaluation of GRU and LSTM architectures for facial expression recognition within an adaptive one-vs-all classification framework, addressing critical gaps in systematic architectural comparison and computational efficiency characterization. The rigorous $2\times2\times2$ factorial experimental design incorporating advanced statistical methods including power analysis, practical equivalence testing, and per-emotion performance assessment establishes methodological standards for architectural comparison studies in facial expression recognition.

The statistical analysis reveals no significant performance differences between GRU and LSTM architectures ($p > 0.05$ across all metrics), with consistently small effect sizes (Cohen's $d \leq 0.306$) suggesting limited practical differences. However, critically inadequate statistical power (capable of detecting only $d \geq 1.43$) necessitates cautious interpretation as preliminary evidence rather than definitive equivalence proof. GRU architectures demonstrate 25% theoretical computational efficiency advantage through simplified gating mechanisms (3 vs 4 gates), translating to reduced matrix operations per timestep. When scaled across the one-vs-all framework comprising six independent binary classifiers, this per-classifier advantage yields substantial cumulative efficiency benefits, making GRU implementations particularly attractive for resource-constrained deployment scenarios.

The proposed adaptive loss function selection mechanism successfully implements systematic loss function assignment based on empirical class imbalance ratios, automatically selecting focal loss (severe imbalance, ratio >11.5), weighted binary cross-entropy (moderate imbalance, ratio 3.5-11.5), or standard binary cross-entropy (balanced scenarios, ratio ≤ 3.5). However, empirical evaluation reveals unexpected underperformance for recall optimization, with standard binary cross-entropy significantly outperforming adaptive approaches ($p=0.002$, $d=-0.787$). This counterintuitive finding suggests that adaptive strategies require refinement through hyperparameter optimization, threshold calibration, and integration with complementary techniques for challenging emotion categories exhibiting severe class imbalance and expression subtlety.

System performance achieves $92.7\% \pm 5.0\%$ overall accuracy on CK+ dataset, demonstrating competitive results compared to state-of-the-art methods while maintaining superior computational efficiency through simplified architectural design. Per-emotion analysis reveals substantial performance heterogeneity (F1-scores: 0.462-0.973), emphasizing the importance of emotion-specific evaluation and targeted optimization strategies enabled by the one-vs-all framework. Practical equivalence testing establishes optimization strategy equivalence for accuracy metrics, supporting utilization of predefined hyperparameter configurations to reduce development overhead across six independent binary classifiers. Optimization strategy equivalence for accuracy metrics supports utilization of predefined hyperparameter configurations, thereby reducing development overhead across six independent binary classifiers.

The statistical analysis framework contributes methodological advances to architectural comparison studies while highlighting fundamental limitations in current experimental designs. The critically inadequate statistical power (observed power ≤ 0.095) emphasizes the imperative for larger-scale validation studies to detect subtle architectural differences reliably across the one-vs-all framework.

Evidence-based recommendations for practical implementation include:

1. **Architectural Selection:** Adopt GRU implementations within one-vs-all frameworks for resource-constrained applications requiring optimal efficiency-performance trade-offs across multiple binary classifiers, pending confirmation through large-scale validation achieving adequate statistical power
2. **Adaptive Loss Function Implementation:** Deploy refined adaptive selection mechanisms with emotion-specific hyperparameter calibration for addressing varying class imbalance severities within independent binary classifiers, incorporating systematic threshold validation and focal loss parameter optimization
3. **Emotion-Specific Optimization:** Implement targeted strategies for challenging emotion categories (fear, sadness) demonstrating F1-scores below 0.50 through specialized data augmentation, architectural modifications, and complementary handling of expression subtlety and intra-class variability
4. **Development Efficiency:** Utilize predefined hyperparameter configurations for accuracy-focused systems to minimize computational overhead during model preparation across the one-vs-all framework, reserving Bayesian optimization for applications requiring optimization of specific performance aspects
5. **Framework Selection:** Consider one-vs-all approaches for applications requiring detailed per-emotion analysis and emotion-specific optimization capabilities, while accounting for $6\times$ training overhead and inference computational requirements in deployment planning

Critical limitations requiring acknowledgment include: critically small sample size restricting detection capabilities for small-to-medium effects across independent binary classifiers, single-dataset validation limiting cross-domain generalizability of adaptive loss function effectiveness, theoretical rather than empirical efficiency validation requiring direct computational measurements within the one-vs-all framework, and adaptive loss

function mechanisms requiring further refinement through extensive empirical validation across diverse imbalance scenarios and expression datasets.

Future investigations should prioritize large-scale validation studies ($n \geq 50$ per condition) achieving adequate statistical power for detecting practically meaningful effects across the one-vs-all framework. Cross-dataset evaluation across multiple facial expression databases (FER2013, AffectNet, RAF-DB) would establish broader generalizability of adaptive loss function selection and architectural efficiency-performance characterizations, while empirical computational efficiency measurements under controlled conditions would provide quantitative deployment guidance for the one-vs-all approach. Systematic investigation of focal loss hyperparameter sensitivity, threshold calibration robustness, and integration with complementary imbalance handling strategies represents essential research directions for enhancing adaptive loss selection effectiveness.

The evidence presented establishes GRU architectures within adaptive one-vs-all frameworks as promising candidates for efficiency-critical facial expression recognition applications, pending confirmation through adequately powered validation studies. The adaptive loss function selection mechanism advances methodological standards for addressing class imbalance challenges in emotion recognition systems, while the comprehensive statistical analysis framework incorporating power analysis, practical equivalence testing, and per-emotion performance assessment provides replicable methodology for future architectural comparison investigations. These contributions provide both preliminary practical guidance and foundational methodology for continued research in computational efficiency optimization and adaptive learning strategies for facial expression recognition systems.

Acknowledgement

This work was supported by the Kemdikbud Research Grant on Penelitian Fundamental - Reguler with grant number 127/C3/DT.05.00/PL/2025 and 028/LL6/PL/AL.04/2025, 118/F.9-05/UDN-09/2025.

REFERENCES

1. P. Radočaj, and G. Martinović, *Emotion Recognition in Autistic Children Through Facial Expressions Using Advanced Deep Learning Architectures*, Applied Sciences, vol. 15, no. 17, p. 9555, August 2025. doi: 10.3390/app15179555.
2. S. S. C. S, and J. R, *Design of a Computational Model to Detect Hybrid Emotion Through Facial Expressions in Videos Using CNN LSTM*, Journal of Mechanics of Continua and Mathematical Sciences, pp. 1984–1993, October 2025. doi: 10.53759/7669/jmc202505155.
3. S. A. Salloum, K. M. Alomari, A. M. Alfaisal, R. A. Aljanada, and A. Basiouni, *Emotion recognition for enhanced learning: using AI to detect students' emotions and adjust teaching methods*, Smart Learning Environments, vol. 12, no. 1, p. 21, February 2025. doi: 10.1186/s40561-025-00374-5.
4. M. Najmabadi, M. Masoudifar, and A. Hajipour, *Weighted classification of deep and traditional histogram-based features with kernel representation for robust facial expression recognition*, Applied Soft Computing, vol. 182, 2025. doi: 10.1016/j.asoc.2025.113630.
5. M. Munsarif, and K. R. Ku-Mahamud, *Deep residual bidirectional long short-term memory fusion: achieving superior accuracy in facial emotion recognition*, Bulletin of Electrical Engineering and Informatics, vol. 14, no. 3, pp. 2143–2155, 2025. doi: 10.11591/eei.v14i3.9090.
6. Q. Yang, Y. He, H. Chen, Y. Wu, and Z. Rao, *A Novel Lightweight Facial Expression Recognition Network Based on Deep Shallow Network Fusion and Attention Mechanism*, Algorithms, vol. 18, no. 8, p. 473, July 2025. doi: 10.3390/a18080473.
7. P. R. Jain, S. M. Khurshid Quadri, and A. Khattar, *PM-ViT a Framework for the Recognition of Emotions and Proclivity toward Mental Illness Using Facial Expressions*, Journal of Computer Science, vol. 21, no. 3, pp. 479–493, 2025. doi: 10.3844/jcssp.2025.479.493.
8. A. R. Dalabehera, S. Bebortta, N. Kumar, and D. Senapati, *Mist–fog-assisted real-time emotion recognition using deep transfer learning framework for smart city 4.0*, Internet of Things, vol. 27, p. 101237, October 2024. doi: 10.1016/j.iot.2024.101237.
9. S. N. Yousafzai, M. Iqbal, H. Khan, M. S. Sarfraz, A. Khalil, M. Adnan, S. Hussain, S. Lee, and J. W. Baek, *A multi-scale simplicial transformer with graph attention for facial emotion recognition*, Ain Shams Engineering Journal, vol. 16, no. 10, p. 103584, October 2025. doi: 10.1016/j.asej.2025.103584.
10. S. Ullah, J. Ou, Y. Xie, and W. Tian, *Facial expression recognition (FER) survey: a vision, architectural elements, and future directions*, PeerJ Computer Science, vol. 10, p. e2024, June 2024. doi: 10.7717/peerj-cs.2024.
11. Y. Zheng, and E. Blasch, *Facial Micro-Expression Recognition Enhanced by Score Fusion and a Hybrid Model from Convolutional LSTM and Vision Transformer*, Sensors, vol. 23, no. 12, p. 5650, June 2023. doi: 10.3390/s23125650.
12. Mujiyanto, A. Setyanto, K. Kusrini, and E. Utami, *Swin Transformer with Enhanced Dropout and Layer-wise Unfreezing for Facial Expression Recognition in Mental Health Detection*, Engineering, Technology and Applied Science Research, vol. 14, no. 6, pp. 19016–19023, 2024. doi: 10.48084/etasr.9139.

13. P. Shi, H. Fang, C. Li, W. Sun, J. Bai, and Y. Zhu, *Uncertain and biased facial expression recognition based on depthwise separable convolutional neural network with embedded attention mechanism*, Journal of Electronic Imaging, vol. 31, no. 4, 2022. doi: 10.1117/1.JEI.31.4.043056.
14. S. K. Sardar, M. C. Cha, and S. C. Lee, *A Comparative Analysis Between Real Human and Virtual Human Interactions in an Academic Learning Context Using Emotion Recognition*, International Journal of Human–Computer Interaction, pp. 1–10, June 2025. doi: 10.1080/10447318.2025.2512526.
15. D. Ciraolo, M. Fazio, R. S. Calabrò, M. Villari, and A. Celesti, *Facial expression recognition based on emotional artificial intelligence for tele-rehabilitation*, Biomedical Signal Processing and Control, vol. 92, p. 106096, June 2024. doi: 10.1016/j.bspc.2024.106096.
16. M. R. Manavand, M. H. Salarifar, M. Ghavami, and M. Taghipour-Gorjikolaie, *Driver's facial expression recognition by using deep local and global features*, Information Sciences, vol. 692, p. 121658, February 2025. doi: 10.1016/j.ins.2024.121658.
17. R. Grover, and S. R. Bansal, *Optimizing Facial Expression Recognition in Challenging Environment: A Streamlined CNN with Pre-processing Techniques*, Journal of The Institution of Engineers (India): Series B, vol. 106, no. 4, pp. 1329–1348, 2025. doi: 10.1007/s40031-024-01184-y.
18. W. J. Baddar, S. Lee, and Y. M. Ro, *On-the-Fly Facial Expression Prediction Using LSTM Encoded Appearance-Suppressed Dynamics*, IEEE Transactions on Affective Computing, vol. 13, no. 1, pp. 159–174, January 2022. doi: 10.1109/TAFFC.2019.2957465.
19. R. Krishna, and K. V. Prema, *Constructing and Optimizing RNN Models to Predict Fruit Rot Disease Incidence in Areca Nut Crop Based on Weather Parameters*, IEEE Access, vol. 11, pp. 110582–110595, 2023. doi: 10.1109/ACCESS.2023.3311477.
20. H. Zhao, *Animation Character Mouth Matching Model Considering Reinforcement Learning and Feature Extraction*, Informatica, vol. 48, no. 3, September 2024. doi: 10.31449/inf.v48i3.6187.
21. P. Majumdar, M. Vatsa, and R. Singh, *Uniform misclassification loss for unbiased model prediction*, Pattern Recognition, vol. 144, p. 109689, December 2023. doi: 10.1016/j.patcog.2023.109689.
22. J. Sun, H. H. Dodge, and M. H. Mahoor, *MC-ViViT: Multi-branch Classifier-ViViT to detect Mild Cognitive Impairment in older adults using facial videos*, Expert Systems with Applications, vol. 238, p. 121929, March 2024. doi: 10.1016/j.eswa.2023.121929.
23. T. Mahbub, A. Obeid, S. Javed, J. Dias, T. Hassan, and N. Werghi, *Center-Focused Affinity Loss for Class Imbalance Histology Image Classification*, IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 2, pp. 952–963, February 2024. doi: 10.1109/JBHI.2023.3336372.
24. T. Dinh, D. Tran, Z. Dobešová, H. V. Hong, D. Lisik, and R. Khan, *An efficient fusion-based deep learning framework for land use and land cover image clustering*, Engineering Applications of Artificial Intelligence, vol. 161, p. 112061, December 2025. doi: 10.1016/j.engappai.2025.112061.
25. V. Tiruvikraman, D. Selvakumar, and P. Vijayakumar, *Real-Time Smart Surveillance and Enforcement for Dust throw Detection and Identity Recognition Using YOLO 12 and SA - FaceXNet*, Signal, Image and Video Processing, vol. 19, no. 12, p. 983, December 2025. doi: 10.1007/s11760-025-04582-x.
26. A. M. Pascual, J. C. Prados, J. M. Ortiz, M. L. López, and I. López, *Light-FER: A Lightweight Facial Emotion Recognition System on Edge Devices*, Sensors, vol. 22, no. 23, p. 9524, December 2022. doi: 10.3390/s22239524.
27. G. Sanil, K. Prakash, S. Prabhu, V. C. Nayak, and S. Sengupta, *2D-3D Facial Image Analysis for Identification of Facial Features Using Machine Learning Algorithms With Hyper-Parameter Optimization for Forensics Applications*, IEEE Access, vol. 11, pp. 82521–82538, 2023. doi: 10.1109/ACCESS.2023.3298443.
28. N. Xie, Z. Liu, Z. Li, W. Pang, and B. Lu, *Student engagement detection in online environment using computer vision and multi-dimensional feature fusion*, Multimedia Systems, vol. 29, no. 6, pp. 3559–3577, December 2023. doi: 10.1007/s00530-023-01153-3.
29. H. Arabian, T. Abdulbaki Alshirbaji, J. G. Chase, and K. Moeller, *Emotion Recognition beyond Pixels: Leveraging Facial Point Landmark Meshes*, Applied Sciences, vol. 14, no. 8, p. 3358, April 2024. doi: 10.3390/app14083358.
30. G. L. Sălăgean, M. Leba, and A. C. Ionica, *Seeing the Unseen: Real-Time Micro-Expression Recognition with Action Units and GPT-Based Reasoning*, Applied Sciences, vol. 15, no. 12, p. 6417, June 2025. doi: 10.3390/app15126417.
31. Shanimol. A, and J. Charles, *ResNet50 and GRU: A Synergistic Model for Accurate Facial Emotion Recognition*, International Journal on Advanced Science, Engineering and Information Technology, vol. 15, no. 8, 2024. doi: 10.14569/IJACSA.2024.0150861.
32. Y. Zhao, W. Zhang, and X. Liu, *Grid search with a weighted error function: Hyper-parameter optimization for financial time series forecasting*, Applied Soft Computing, vol. 154, p. 111362, March 2024. doi: 10.1016/j.asoc.2024.111362.
33. F. Ehsani, and M. Hosseini, *Customer churn prediction using a novel meta-classifier: an investigation on transaction, Telecommunication and customer churn datasets*, Journal of Combinatorial Optimization, vol. 48, no. 1, p. 7, August 2024. doi: 10.1007/s10878-024-01196-w.
34. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. 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, San Francisco, CA, USA, pp. 94–101, June 2010. doi: 10.1109/CVPRW.2010.5543262.
35. C. Bisogni, L. Cimmino, M. De Marsico, F. Hao, and F. Narducci, *Emotion recognition at a distance: The robustness of machine learning based on hand-crafted facial features vs deep learning models*, Image and Vision Computing, vol. 136, p. 104724, August 2023. doi: 10.1016/j.imavis.2023.104724.
36. M. Mohana, P. Subashini, and M. Krishnaveni, *Emotion Recognition from Facial Expression Using Hybrid CNN–LSTM Network*, International Journal of Pattern Recognition and Artificial Intelligence, vol. 37, no. 08, p. 2356008, June 2023. doi: 10.1142/S0218001423560086.
37. N. Kumar H N, A. S. Kumar, G. Prasad M S, and M. A. Shah, *Automatic facial expression recognition combining texture and shape features from prominent facial regions*, IET Image Processing, vol. 17, no. 4, pp. 1111–1125, March 2023. doi: 10.1049/ipr2.12700.
38. Bakiaraj M, Subramani B *Optimized hybrid deep learning pipelines for processing heterogeneous facial expression datasets*, Measurement: Sensors, vol. 31, pp. 100938, February 2024. doi: 10.1016/j.measen.2023.100938.
39. Lu nannan, Tan Zhen, Qian Jiansheng *MRSLN: A Multimodal Residual Speaker-LSTM Network to alleviate the over-smoothing issue for Emotion Recognition in Conversation* Neurocomputing, vol. 580, pp. 127467, May 2024. 10.1016/j.neucom.2024.127467

40. J. Zhu, B. Luo, T. Yang, Z. Wang, X. Zhao and Y. Ga *Knowledge Conditioned Variational Learning for One-Class Facial Expression Recognition* IEEE Transactions on Image Processing, vol. 32, pp. 4010-4023, 2023 10.1109/TIP.2023.3293775