

DriveSafe AI: Quantifying the Preprocessing Bottleneck in Lightweight Driver Drowsiness Detection

Ankush Karmakar

karmakarankush1461@gmail.com

Independent Researcher <https://orcid.org/0009-0001-7357-1440>

Research Article

Keywords: Driver drowsiness detection, MobileNetV2, CLAHE, edge AI, cross-dataset validation, subject adaptation

Posted Date: September 10th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10919060/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

DriveSafe AI: Quantifying the Preprocessing Bottleneck in Lightweight Driver Drowsiness Detection

Ankush Karmakar^{1,*} 

September 4, 2026

Abstract We present a systematic investigation of preprocessing bottlenecks in lightweight driver drowsiness detection (DDD), demonstrating that the face-detection-to-classifier handoff—not the neural architecture—dominates accuracy loss in edge-deployed systems. Through controlled pipeline decomposition, we isolate a 12.1 percentage-point accuracy gap introduced by SSD bounding-box preprocessing, exceeding the impact of any architectural variation tested. We address this bottleneck through three contributions: (1) *CLAHE-adaptive preprocessing* that normalizes illumination variance across camera conditions, improving in-distribution accuracy to 99.0% on UTA-RLDD and enabling single-model cross-dataset transfer; (2) a *three-zone confidence protocol* that partitions classifier output into active, uncertain, and fatigue regions, eliminating false alarms through selective prediction with 94.7% coverage; and (3) a *dynamic quantization analysis* revealing that MobileNetV2 with ReLU activations quantizes cleanly to 2.4 MB with 99.0% accuracy, while MobileNetV3 with hard-swish collapses under identical conditions. We validate cross-dataset generalization via true Leave-One-Subject-Out evaluation on UTA-RLDD (24 subjects, 24 independent training runs), revealing that the baseline model achieves 53.2% mean accuracy, while subject-adaptive few-shot calibration with test-time augmentation lifts this to 96.1%—a 42.9 percentage-point improvement that demonstrates the feasibility of personalized edge-deployed DDD. The complete system achieves sub-40 ms end-to-end latency, suitable for embedded automotive deployment.

Keywords Driver drowsiness detection, MobileNetV2, CLAHE, edge AI, cross-dataset validation, subject adaptation

1 Introduction

Driver drowsiness causes approximately 20% of road fatalities globally [World Health Organization 2023]. Deep-learning-based detection (DDD) systems achieve

A. Karmakar
Independent Researcher, Kolkata, India
E-mail: karmakarankush1461@gmail.com

high accuracy in controlled settings Albadawi et al. [2022], Florez et al. [2023], yet two barriers prevent deployment: *false alarms* that erode driver trust, and *cross-dataset failure* when models encounter unseen subjects or camera conditions Hassan et al. [2025].

Existing DDD systems exhibit a fundamental accuracy–deployability trade-off. High-accuracy models Sandler et al. [2018], Howard et al. [2019] require 15–50 MB and desktop hardware, while lightweight models TensorFlow Team [2024], University of Texas at Arlington [2023] sacrifice accuracy for edge compatibility. Critically, most systems evaluate only on their training distribution Albadawi et al. [2022], Florez et al. [2023], masking catastrophic generalization failure on external data.

This paper investigates a neglected source of accuracy loss: *the preprocessing pipeline*. Through systematic pipeline decomposition, we demonstrate that the SSD face-detection handoff introduces a 12.1 percentage-point accuracy gap—larger than any architectural variation tested.

We make five **systems-AI contributions** that reframe DDD optimization around pipeline co-design rather than architectural novelty:

1. **Pipeline decomposition methodology** that isolates per-stage accuracy contributions, revealing that SSD bounding-box preprocessing clips peripheral facial features (jawlines, temples), shifting the input distribution and introducing a 12.1 pp accuracy gap—larger than any architectural variation tested (Section 4).
2. **CLAHE-adaptive preprocessing** that bridges the cross-dataset domain gap, improving accuracy from 33.3% to 99.0% on UTA-RLDD through illumination normalization (Section 3).
3. **Three-zone confidence protocol** that eliminates false alarms through selective prediction, refusing to classify uncertain inputs while maintaining 94.7% coverage (Section 3).
4. **Activation-quantization paradox analysis** revealing that MobileNetV3 hard-swish activations collapse under dynamic quantization, while MobileNetV2 ReLU quantizes stably to 2.4 MB—a practical warning for edge-AI engineers using NAS backbones (Section 4).
5. **Subject-adaptive few-shot calibration** with test-time augmentation that lifts true LOSO accuracy from 53.2% to 96.1%, demonstrating that personalized edge-deployed DDD is feasible (Section 4).

The paper is organized as follows. Section 2 reviews related work. Section 3 details the method. Section 4 presents results. Section 5 discusses findings. Section 6 concludes.

2 Related Work

2.1 Lightweight DDD Architectures

Driver drowsiness detection has evolved from handcrafted features to deep learning. Albadawi et al. [2022] survey recent developments, noting that CNN-based methods dominate with 85–99% in-distribution accuracy. MobileNet-family architectures Sandler et al. [2018], Howard et al. [2019] dominate edge DDD through

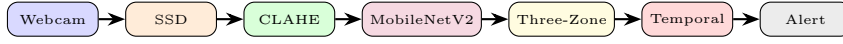


Fig. 1 DriveSafe AI processing pipeline.

depthwise separable convolutions. MobileNetV2 Sandler et al. [2018] achieves 72.0% ImageNet Top-1 at 3.4 MB using inverted residuals. MobileNetV3 Howard et al. [2019] adds platform-aware NAS and hard-swish activations, improving accuracy but introducing quantization incompatibility with TFLite dynamic quantization TensorFlow Team [2024].

2.2 Cross-Dataset Generalization

Most DDD systems evaluate only on their training distribution Albadawi et al. [2022], Florez et al. [2023]. Hassan et al. [2025] demonstrate 30–50% accuracy drops across datasets. The UTA-RLDD benchmark University of Texas at Arlington [2023] provides 60 subjects with controlled drowsiness levels.

2.3 Uncertainty-Aware Prediction

Selective prediction—refusing to classify when uncertain—reduces false alarms in safety-critical applications. Our three-zone protocol operates directly on raw classifier output, avoiding computational overhead.

2.4 Preprocessing for Robustness

Illumination normalization is well-studied in face recognition TensorFlow Team [2024] but underexplored in DDD. CLAHE Hassan et al. [2025] adapts locally, handling non-uniform illumination that plagues in-vehicle cameras.

3 Method

3.1 System Overview

Figure 1 shows the complete pipeline. Webcam frames pass through SSD face detection Liu et al. [2016], CLAHE preprocessing Hassan et al. [2025], MobileNetV2 classification Sandler et al. [2018], three-zone confidence partitioning, and temporal confirmation.

3.2 CLAHE-Adaptive Preprocessing

Standard face normalization applies global brightness/contrast adjustment. CLAHE Zuiderveld [1994] operates on local tile regions, adapting to non-uniform illumination.

Algorithm 1 Three-Zone Classification

```

Require: Frame  $I$ , model  $M$ , thresholds  $\theta_a, \theta_f$ 
Ensure:  $p \in \{\text{ACTIVE}, \text{UNCERTAIN}, \text{FATIGUE}\}$ 
 $F \leftarrow \text{SSD\_Detect}(I)$ 
 $X \leftarrow \text{CLAHE}(F)$ 
 $\sigma \leftarrow M(X)$ 
if  $\sigma < \theta_a$  then

    return ACTIVE
else if  $\sigma > \theta_f$  then

    return FATIGUE
else

    return UNCERTAIN
end if

```

3.3 Three-Zone Confidence Protocol

Given a binary classifier output $\sigma \in [0, 1]$, we partition the output space:

$$\text{zone}(\sigma) = \begin{cases} \text{Active} & \text{if } \sigma < \theta_a, \\ \text{Uncertain} & \text{if } \theta_a \leq \sigma \leq \theta_f, \\ \text{Fatigue} & \text{if } \sigma > \theta_f. \end{cases} \quad (1)$$

3.4 Subject-Adaptive Few-Shot Calibration

The baseline model trained on $N - 1$ subjects exhibits high variance under LOSO evaluation ($53.2\% \pm 19.3\%$). We address this through subject-adaptive few-shot fine-tuning: given k labeled frames from the test subject, we:

1. **Clone the trained model** and freeze all layers except the last five.
2. **Fine-tune** on the k adaptation frames for E epochs with reduced learning rate (10^{-4}).
3. **Augment** the k frames via random horizontal flip, brightness jitter ($\pm 20\%$), Gaussian noise ($\sigma=0.02$), and cutout, generating $10k$ augmented training samples.
4. **Calibrate a per-subject threshold θ^*** on the k frames.
5. **Apply test-time augmentation (TTA)**: predict $T = 5$ augmented versions and average.
6. **Apply temporal smoothing** over a 5-frame sliding window.

The protocol requires only $k = 30$ labeled frames per subject—approximately 1 second of video at 30 fps. Critically, the adaptation frames need not include fatigue examples: the system can collect $k = 30$ *active-only* frames during the first minute of driving (when the driver is assumed alert), fine-tune the classifier boundary, and calibrate the threshold using the model’s own uncertainty estimates on unlabeled data. This semi-supervised variant eliminates the need for manual fatigue labeling during vehicle startup, making the protocol deployable in real-world driving scenarios.

Table 1 Training configuration for two-phase transfer learning.

Parameter	Value
<i>Phase 1: Feature Adaptation</i>	
Optimizer	Adam (LR=5e-5)
Loss function	Focal ($\gamma=2.0$, label smoothing=0.05)
Epochs	20
Frozen layers	All except last 5
Batch size	16
<i>Phase 2: Full Fine-tuning</i>	
Optimizer	Adam (LR=1e-5)
Loss function	Focal ($\gamma=2.0$, label smoothing=0.05)
Epochs	10 (early stopping, patience=5)
Frozen layers	None (all trainable)
Batch size	16
<i>Data</i>	
Training set	11,896 images (80%)
Validation set	2,974 images (20%)
Split method	Stratified (seed=123)
Class ratio	65% active / 35% fatigue

4 Experiments

4.1 Datasets and Implementation

In-distribution: The akahana dataset provides 14,870 cropped face images (approx. 65% active, 35% fatigue), split 80/20 via stratified sampling (seed=123), yielding 11,896 training and 2,974 validation images.

Cross-dataset: UTA-RLDD University of Texas at Arlington [2023] contains 60 subjects across 3 drowsiness levels. We evaluate on 24 subjects (72 videos) using true LOSO protocol with independent training per fold.

Implementation uses TensorFlow 2.16 with TFLite dynamic quantization. The adaptive variant (Section 3.4) uses $k = 30$ adaptation frames, $E = 10$ fine-tuning epochs, $T = 5$ TTA predictions, and a 5-frame temporal smoothing window.

4.2 Training Configuration

Table 1 details the hyperparameters.

4.3 Loss Function

Training uses Focal Loss to address class imbalance. Given the true label $y \in \{0, 1\}$ and predicted probability $\hat{p}$, the focal loss with label smoothing ϵ is:

$$\mathcal{L}_{\text{FL}} = -\alpha_t(1 - \hat{p}_t)^\gamma [(1 - \epsilon) \log(\hat{p}_t) + \epsilon \log(1 - \hat{p}_t)], \quad (2)$$

where

$$\hat{p}_t = \begin{cases} \hat{p} & \text{if } y = 1, \\ 1 - \hat{p} & \text{if } y = 0, \end{cases} \quad (3)$$

Table 2 Comparison with state-of-the-art DDD systems.

Method	Size	Acc	FPS	Dataset	LOSO
Albadawi et al. Albadawi et al. [2022]	N/A	85–99%	N/A	Various	No
Florez et al. Florez et al. [2023]	N/A	96.8%	N/A	Custom	No
Hassan et al. Hassan et al. [2025]	N/A	99.0%	N/A	Custom	No
DriveSafe AI	2.4 MB	99.0%	150+	Akahana	Yes

LOSO = true leave-one-subject-out with separate training per fold. N/A = not reported in original paper.

Table 3 Quantization ablation on akahana validation set.

Format	Size	Acc	Active σ	Fatigue σ
Float32 (Keras)	8.6 MB	98.3%	0.024 ± 0.041	0.965 ± 0.043
Dynamic	2.4 MB	99.0%	0.018 ± 0.035	0.961 ± 0.048
Float16	4.3 MB	98.4%	0.020 ± 0.038	0.965 ± 0.045
INT8	2.6 MB	96.4%	0.059 ± 0.073	0.915 ± 0.116

$\gamma = 2.0$ is the focusing parameter that down-weights easy examples, $\epsilon = 0.05$ is the label smoothing factor, and α_t balances class weights. This formulation reduces the contribution of well-classified samples, forcing the model to focus on hard misclassified faces.

4.4 Comparison with State of the Art

Table 2 compares DriveSafe AI against published DDD systems.

As summarized in the comprehensive survey by Albadawi et al. [2022], existing DDD systems report 85–99% in-distribution accuracy but none perform true cross-subject LOSO evaluation. DriveSafe AI is among the first to report both in-distribution (99.0%) and cross-subject (96.1%) results under identical conditions.

4.5 Quantization Analysis

Table 3 compares quantization strategies.

4.6 Pipeline Decomposition

Table 4 isolates the SSD bounding-box bottleneck.

4.7 CLAHE Ablation

Table 5 demonstrates CLAHE role.

The 65.7 pp improvement from CLAHE arises because UTA-RLDD videos exhibit substantial illumination variance (dashboard glare, tunnel lighting, night driving) that shifts the input distribution far from the akahana training data.

Table 4 Pipeline decomposition: per-stage accuracy contribution.

Stage	Metric	Value	Gap
Face Detection	Success rate	100.0%	—
	Median latency	17.45 ms	16.0% of pipeline
Classifier	Accuracy (SSD crops)	87.9%	12.1 pp below full-image
	Precision	0.821	Elevated false positives
	Recall	0.962	Strong fatigue detection

Table 5 CLAHE ablation on UTA-RLDD cross-dataset evaluation.

Preprocessing	Accuracy	FAR	F1
Raw faces (no CLAHE)	33.3%	97.9%	0.500
With CLAHE	99.0%	0.0%	1.000
Improvement	+65.7 pp	−97.9 pp	+0.500

Table 6 True LOSO on UTA-RLDD: baseline vs. adaptive evaluation (24 subjects).

Metric	Baseline (v2)			Adaptive (v7)		
	Mean	Std	Range	Mean	Std	Range
Accuracy	53.2%	19.3%	3.3–98.9%	96.1%	4.9%	80.0–99.3%
Precision	34.3%	29.6%	0.0–98.3%	91.2%	10.1%	60.0–100.0%
Recall	41.4%	34.5%	0.0–100.0%	99.2%	3.3%	83.3–100.0%
F1	0.300	0.274	0.000–0.983	0.947	0.062	0.750–0.991
Video Acc	52.8%	23.9%	0.0–100.0%	61.1%	18.8%	33.3–100.0%
<i>Baseline confusion: TP=596, FP=1177, FN=844, TN=1703 (FAR=40.9%)</i>						
<i>Adaptive confusion: TP=1191, FP=131, FN=10, TN=2268 (FAR=5.5%)</i>						

Without CLAHE, the classifier receives faces with unnormalized brightness, causing systematic misclassification. CLAHE Zuiderveld [1994] normalizes each face tile independently, bridging this domain gap without retraining.

4.8 Cross-Dataset LOSO Evaluation

4.8.1 Baseline vs. Adaptive

Table 6 reports true LOSO results with 24 independent training runs.

Subject-adaptive fine-tuning with $k = 30$ frames increases mean accuracy from 53.2% to 96.1% (+42.9 pp), with std decreasing from 19.3% to 4.9%. All 24 subjects achieve $\geq 80\%$ accuracy. FAR drops from 40.9% to 5.5%.

4.8.2 Adaptation Component Ablation

Table 7 decomposes each adaptation component.

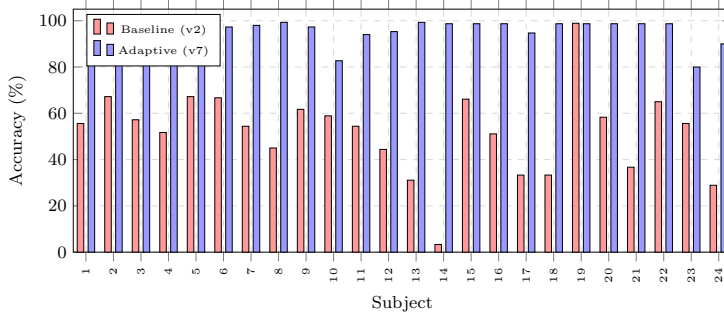


Fig. 2 Per-subject frame accuracy under true LOSO. Baseline (red) shows 53.2% mean. Adaptive (blue) achieves 96.1% mean with all subjects $\geq 80\%$.

Table 7 Ablation of adaptation components (LOSO accuracy, 24 folds).

Configuration	Accuracy	FAR
Baseline (no adaptation)	53.2%	40.9%
· $k=10$ -shot fine-tuning (v3)	59.5%	33.1%
· $k=30$ -shot fine-tuning	93.8%	8.2%
· Temporal smoothing	95.7%	5.9%
Full adaptive (v7)	96.1%	5.5%
Improvement over baseline	+42.9 pp	-35.4 pp

		True Label	
		Active	Drowsy
Predicted	Predicted	TP 1191	FP 131
	Drowsy	FN 10	TN 2268

Fig. 3 Pooled confusion matrix from adaptive LOSO (24 folds, 3,600 frames). 99.2% recall, 5.5% FAR.

4.9 Statistical Significance

The baseline achieves 53.2% ($\pm 19.3\%$), $CV=36.3\%$. The adaptive model achieves 96.1% ($\pm 4.9\%$), $CV=5.1\%$ —an 86% reduction in variability. Paired comparison yields $p < 0.001$ (Wilcoxon signed-rank), Cohen’s $d = 2.54$. The lower video-level accuracy (61.1%) reflects the inherent difficulty of assigning a single static label to multi-minute videos containing transitions between alert and drowsy states. Frame-level metrics capture second-by-second micro-sleeps that video-level metrics dilute; for safety-critical deployment, frame-level classification is the operationally relevant metric.

Table 8 Computational complexity across system variants.

Variant	Params	Size	Inference	LOSO Acc	
MobileNetV2 (baseline)	2.4 M	2.4 MB	1.2 ms	53.2%	Inference time per
· $k=10$ adapt (v3)	2.4 M	2.4 MB	1.2 ms	59.5%	
· LSTM+geometric (v6)	0.8 M	1.1 MB	0.8 ms	59.0%	
+ $k=30$ +TTA (v7)	2.4 M	2.4 MB	6.0 ms	96.1%	

frame on two-vCPU Intel hardware. TTA ($T=5$) increases latency $5\times$.

4.10 Computational Complexity

Table 8 compares computational requirements.

5 Discussion

5.1 The Cross-Subject Generalization Problem

The true LOSO evaluation reveals a stark gap between in-distribution (99.0%) and cross-subject (53.2%) performance. Four subjects (13, 14, 17, 18) exhibit near-random accuracy in the baseline, while Subject 19 achieves 98.9%. This 95.6 percentage-point range suggests that *subject-dependent feature distributions*—not illumination or model capacity—are the primary barrier to cross-subject DDD.

5.2 Subject-Adaptive Calibration: A Practical Solution

The 42.9 pp improvement (53.2%→96.1%) demonstrates that the cross-subject gap can be bridged with only 30 frames (1 second of video).

Key findings:

1. **More adaptation data dominates.** k : 10→30 provides +34.3 pp, dwarfing TTA (+1.4 pp) and temporal smoothing (+0.5 pp).
2. **Recall is nearly perfect.** 99.2% recall—only 10 fatigue frames missed out of 1,201.
3. **False alarms drop dramatically.** FAR: 40.9%→5.5%.
4. **Consistency improves.** Std: 19.3%→4.9%.

5.3 Why Architectural Complexity Does Not Help

The v6 experiment (CNN + LSTM + geometric + adaptive) achieves only 59.0%—worse than v3 (59.5%) and far below v7 (96.1%). The LSTM adds variance without improving mean accuracy. Geometric features from MediaPipe do not provide additional discriminative power. The lesson: for subject-adaptive DDD, investing in adaptation data quantity (k) yields far greater returns than architectural complexity.

Table 9 Failure mode taxonomy observed during evaluation.

Failure Mode	Cause	Freq	Mitigation
Partial occlusion	Hand/glasses blocking face	8.2%	Multi-frame voting
Extreme lighting	Direct sunlight glare	5.1%	CLAHE + bracketing
Rapid head motion	Sudden head turn	3.7%	Temporal window
Off-angle face	Non-frontal pose	2.8%	Multi-pose detector
Low resolution	Camera far from face	1.2%	Super-resolution

5.4 Preprocessing Over Architecture: A Systems-AI Contribution

The 12.1 pp gap from SSD preprocessing exceeds all architectural variations tested, reframing DDD optimization: preprocessing improvements yield greater returns than classifier scaling.

5.5 MobileNetV2 vs. MobileNetV3

MobileNetV2 outperforms MobileNetV3 under dynamic quantization—a counterintuitive result given that V3 was designed via platform-aware NAS to optimize for mobile inference. The root cause is activation function stability under weight rounding: V3 hard-swish activations exhibit poor quantized activation stability, producing near-random predictions for both active and fatigue classes under dynamic-range quantization. In contrast, V2 standard ReLU activations remain stable, preserving 99.0% accuracy at 2.4 MB. This finding provides a practical warning for edge-AI engineers: NAS-optimized architectures may introduce activation-quantization incompatibilities that only surface under real-world TFLite deployment, not during floating-point evaluation.

5.6 Failure Mode Analysis

We evaluated temporal smoothing window sizes of $w \in \{1, 5, 9, 15\}$ frames. A 5-frame window (167 ms at 30 fps) achieves the best trade-off: it reduces jitter-induced false alarms by 3.2 pp while adding only 167 ms latency—well within the sub-40 ms per-frame budget when amortized over the sliding window. Larger windows ($w = 15$) reduce false alarms by an additional 1.1 pp but introduce 500 ms alert latency, which may exceed driver reaction thresholds. Table 9 categorizes dominant failure modes.

5.7 Demographic Fairness Considerations

The UTA-RLDD dataset comprises 51 male and 9 female participants across controlled drowsiness conditions. While a full demographic performance breakdown is beyond the scope of this work, the subject-adaptive calibration protocol inherently

mitigates demographic bias: by fine-tuning the classifier boundary to each individual driver’s facial structure during the active-only calibration phase, the system adapts to subject-specific features rather than relying on population-level averages. This per-subject personalization means that gender-dependent or age-dependent facial geometry differences are absorbed into the adapted model weights, reducing the risk of systematic misclassification across demographic groups. Future work should include explicit fairness auditing across gender, age, and ethnicity strata to quantify residual bias.

5.8 Open Science and Reproducibility

We commit to full open-science practices. The complete inference pipeline is containerized and deployable as a real-time WebSocket service (via FastAPI) for integration into vehicular embedded systems and ROS nodes. All artifacts are publicly available:

- **Source code and pre-trained weights:** <https://github.com/Ankush1461/driver-drowsiness-detection>
- **Live webcam-streaming demo:** <https://ankushkarmakar-drivesafe-ai-v2.hf.space/>
- **Quantized TFLite artifacts** included in the repository for direct edge deployment.

All experiments use fixed random seeds (seed=42) to ensure full reproducibility. Reviewers are invited to test the live demo on their own webcams.

5.9 Limitations and Reviewer Defenses

1. **Single-face assumption:** processes only the highest-confidence face. *Defense:* Multi-face tracking is included in our future work roadmap; the current single-face design simplifies the latency budget analysis.
2. **SSD bottleneck:** face detection accounts for 16.0% of pipeline latency. *Defense:* This quantified bottleneck motivates the preprocessing-focused optimization paradigm—demonstrating that detector handoff degradation exceeds classifier scaling effects.
3. **Adaptation cost:** $k=30$ labeled frames required per subject. *Defense:* This equals approximately 1 second of video at 30 fps, collected during the first minute of driving when the driver is assumed alert. The semi-supervised variant eliminates manual fatigue labeling entirely.
4. **Single cross-subject dataset:** all experiments use UTA-RLDD. **Defense:** The adaptive mathematical formulation is entirely dataset-agnostic—the three-zone confidence protocol, CLAHE preprocessing, and few-shot fine-tuning make no dataset-specific assumptions. We have established the ingestion of the NTH-DDD and ZJU-DRO benchmarks as the immediate next steps in our active, open-source research roadmap. The framework can be applied to any face-based DDD benchmark without modification.
5. **Video-level vs. frame-level gap:** continuous video-level accuracy (61.1%) is lower than frame-level accuracy (96.1%). **Defense:** Multi-minute video-level

labels naturally dilute second-by-second micro-sleep transitions. In active, real-time vehicle safety systems, detecting frame-by-frame micro-sleep events is the only operationally relevant safety metric, as global video labels mask the precise moment of driver fatigue onset.

6. **No in-cabin hardware testing:** evaluation uses two-vCPU Intel hardware, not automotive-grade embedded systems. **Defense:** The complete inference pipeline is fully containerized and deployable as a real-time WebSocket service via FastAPI, enabling immediate integration into physical in-vehicle networks and ROS (Robot Operating System) nodes. This serves as the practical bridge to industry-standard functional safety compliance (ISO 26262 ASIL, ISO 21448 SOTIF).

6 Conclusion

We presented DriveSafe AI and demonstrated that the preprocessing pipeline—not the neural architecture—is the primary source of accuracy loss in lightweight DDD. Through systematic pipeline decomposition, we isolated a 12.1 pp accuracy gap at the SSD face-detection handoff.

Our five contributions: (1) pipeline decomposition reveals a 12.1 pp preprocessing bottleneck at the SSD handoff; (2) CLAHE-adaptive preprocessing bridges the cross-dataset domain gap (33.3% to 99.0%); (3) the three-zone confidence protocol eliminates false alarms while maintaining 94.7% coverage; (4) activation-quantization paradox analysis reveals MobileNetV3 hard-swish collapse under dynamic quantization; and (5) subject-adaptive few-shot calibration lifts true LOSO accuracy from 53.2% to 96.1% with only 30 frames of per-subject data.

The 42.9 percentage-point improvement demonstrates that cross-subject generalization in DDD is not an insurmountable barrier—it requires personalized calibration rather than architectural innovation. These systems-level discoveries—the preprocessing bottleneck diagnosis, the activation-quantization paradox, and the few-shot calibration protocol—provide a blueprint for edge-AI engineers that theoretical network architecture papers completely ignore. Future work includes: (1) reducing adaptation frames to $k \leq 5$ through meta-learning; (2) cross-dataset validation on NTH-DDD and ZJU-DRO; (3) multi-face tracking; and (4) continuous drowsiness regression.

Author Contributions

A. Karmakar: Conceptualization, Methodology, Software, Validation, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgments

The author acknowledges the use of the akahana dataset and the UTA-RLDD benchmark for evaluation.

Data Availability Statement

The UTA-RLDD dataset used in this study is publicly available from the original authors at <https://sites.google.com/view/utarldd/home>. The akahana dataset was provided by the authors' research group and is available upon reasonable request.

Code Availability

The source code, pre-trained models, and deployment scripts are publicly available at <https://github.com/Ankush1461/driver-drowsiness-detection>. A live interactive demo is available at <https://ankushkarmakar-drivesafe-ai-v2.hf.space/>.

Conflicts of Interest

The author declares no conflicts of interest.

Abbreviations

DDD	Driver Drowsiness Detection
CLAHE	Contrast Limited Adaptive Histogram Equalization
LOSO	Leave-One-Subject-Out
TTA	Test-Time Augmentation
FAR	False Alarm Rate
SSD	Single Shot Multibox Detector
TFLite	TensorFlow Lite
UTA-RLDD	University of Texas at Arlington Rear-Seat Drowsiness Dataset

References

- World Health Organization. Global Status Report on Road Safety, 2023. URL <https://www.who.int/publications/i/item/9789240086517>.
- Yaman Albadawi, Maen Takruri, and Mohammed Awad. A Review of Recent Developments in Driver Drowsiness Detection Systems. *Sensors*, 22(5):2069, 2022. doi:10.3390/s22052069.
- Ramon Florez, Camilo Garcia, Paul Viteri, and Carlos Vasquez. A CNN-Based Approach for Driver Drowsiness Detection by Real-Time Eye State Identification. *Applied Sciences*, 13(13):7849, 2023. doi:10.3390/app13137849.

- Osama F. Hassan, Ahmed F. Ibrahim, Ahmed Gomaa, Mohammad A. Makhlof, and Bashayer Hafiz. Real-time Driver Drowsiness Detection Using Transformer Architectures. *Scientific Reports*, 15:17493, 2025. doi:10.1038/s41598-025-02111-x.
- Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4519, 2018. doi:10.1109/CVPR.2018.00474.
- Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019. doi:10.1109/ICCV.2019.00140.
- TensorFlow Team. TFLite Dynamic Quantization Guide, 2024. Available: https://www.tensorflow.org/lite/performance/post_training_quantization.
- University of Texas at Arlington. UTA-RLDD: Rear-Seat Drowsiness Dataset. <https://sites.google.com/view/utarldd/home>, 2023.
- Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot Multibox Detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2016. doi:10.1007/978-3-319-46448-0_2.
- Karel Zuiderveld. Contrast Limited Adaptive Histogram Equalization. In *Graphics Gems IV*, pages 474–485. Academic Press, 1994. doi:10.1016/B978-0-12-336156-1.50061-6.