Attention-Guided Lightweight MobileNetV2 for Real-Time Driver Drowsiness Classification on Edge-IoT Systems

Yo Ceng Giap[1], Muljono[2]*, Affandy[3], Ruri Suko Basuki[4],
Harun Al Azies[5], R. Rizal Isnanto[6], Deshinta Arrova Dewi[7]
Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia [1, 2, 3, 4, 5]
Faculty of Science and Technology, Buddhi Dharma University, Banten, Indonesia [1]
Doctoral Program of Information Systems-School of Postgraduate Studies, Diponegoro University, Indonesia[6]
Faculty of Data Science and Information Technology, INTI International University, Malaysia [7]

***Abstract*—Driver drowsiness is a major cause of traffic accidents, so Edge-IoT platforms with limited resources need to be able to accurately and quickly detect when drivers are drowsy. This study examines attention-guided lightweight CNN design predicated on MobileNetV2 for real-time driver drowsiness detection. The authors compare a SE-enhanced MobileNetV2 to the baseline model and a structurally optimized version that uses Depthwise Separable Convolution (DSC), Bottleneck blocks, and Expansion layers. Experiments on 500 images demonstrate that channel attention enhances feature discrimination, whereas structural optimization yields the most resilient trade-off between accuracy and latency. Statistical validation employing 95% confidence intervals and two-proportion Z-tests substantiates the significance of these enhancements. The proposed models support real-time inference despite their small size (about 2.6 million parameters and 315 million FLOPs). These findings suggest structural optimization is more important than attention mechanisms in designing lightweight CNNs for embedded driver monitoring.**

***Keywords*—Driver drowsiness detection; Edge-IoT deployment; lightweight convolutional neural networks; process innovation; MobileNetV2 optimization; squeeze-and-excitation attention**

I. Introduction

Drowsy driving remains one of the leading causes of car accidents worldwide. This is mostly because it slows reaction time, reduces alertness, and makes decision-making harder [1]. Recent research underscores the substantial influence of fatigue on traffic incidents and stresses the necessity for automated, vision-based detection systems proficient in recognizing ocular signs of diminished alertness [2]. Vision-based driver monitoring has emerged as a predominant research focus, as it facilitates the identification of fatigue-related facial behaviors—such as blinking patterns, eyelid closure, yawning, and head movements—that are significantly associated with the progression of drowsiness [3] [4].

As more and more cars adopt artificial intelligence, there is a growing need for lightweight, real-time, and reliable fatigue-detection models that can run effectively on Edge-IoT platforms with limited computing power [5]. Recent real-time applications, such as MobileNet-based drowsiness detectors, further demonstrate the utility and efficiency of lightweight architectures for embedded deployment [5].

Deep learning has significantly advanced driver-state analysis by enabling image-based models to learn complex, subtle visual patterns that are difficult to capture with handcrafted features [6]. High-capacity CNN and transformer-based models, such as EffRes-DrowsyNet and MG-YOLOv8, exhibit robust benchmark performance [3], yet their computational complexity restricts deployment on embedded devices. Recent studies in lightweight visual recognition emphasize diminishing model complexity via depthwise separable convolutions, inverted bottlenecks, and fast feature recalibration [7], with hybrid attention architectures that selectively amplify relevant features without significant processing burden [8]. Studies such as SCAT, which integrates spatial and channel-enhanced self-attention, show that combining local and global feature representations significantly improves accuracy while maintaining efficiency—demonstrating the importance of attention-guided lightweight designs [9].

Nguyen et al. showed that DSC with dilated filters makes embedded inference on limited hardware more efficient [10]. Lightweight CNNs can perform well when used to support users in real-time fatigue detection, even with limited training data in driver monitoring applications [11]. Nonetheless, despite their efficacy, lightweight CNNs sometimes encounter difficulties in detecting nuanced fatigue-related micro-expressions—such as partial eyelid closure and micro-blinks—which are essential for the early identification of drowsiness in practical driving situations. To address this problem, attention mechanisms such as SE, CBAM, ECA, and their hybrid variants have been shown to effectively improve channel selectivity and spatial focus in vision tasks. Hybrid attention, especially, has led to big improvements in fine-grained classification tasks with small texture changes. For example, hybrid-attention Xception models for brain tumor analysis [12] show that lightweight architectures gain significant benefit from attention-based improvements. SE is well-suited to lightweight architectures because it adds channel recalibration with very little extra computational cost, unlike spatial attention or hybrid multi-branch attention mechanisms.

*Corresponding author.

Even with these improvements, the authors still don't fully understand how attention mechanisms work in lightweight CNN-based driver drowsiness classification, especially when compared directly to structural lightweight optimizations. There has not been a systematic, statistically validated study to determine whether channel attention mechanisms, such as Squeeze-and-Excitation, offer benefits similar to or complementary to architectural changes in Edge-IoT settings, especially for MobileNetV2, a popular backbone for environments with limited resources.

This research introduces an Attention-Guided Lightweight MobileNetV2 to facilitate real-time driver drowsiness classification on Edge-IoT systems, addressing the identified gap. More importantly, the study conducts a controlled, statistically sound comparison of attention-based refinement and structural lightweight optimization within the same MobileNetV2 backbone. This framework offers empirical insights into the impact of various lightweight enhancement strategies on fine-grained fatigue-feature recognition, computational efficiency, and deployment suitability in resource-constrained environments.

II. Related Work

Driver drowsiness detection has been widely explored across three primary categories: facial-feature–based methods, physiological-signal–based approaches, and vehicle-behavior analysis [13], [14], [15]. Vision-based facial-feature methods are still the most useful and widely used of these. They use cues like how often someone blinks, how their eyelids close, how often they yawn, and how their head moves [3], [5], [2]. But classical methods often don't perform as well when lighting changes, there are obstructions, or the appearance of different drivers varies, which is why the authors need to use strong deep learning models that can work in real driving situations [16] [17].

Deep learning–based methods have substantially improved fatigue detection by learning hierarchical facial representations [18]. Lightweight CNNs such as MobileNet, DenseNet, ResNet50V2, and VGG19 have been applied to eye-state classification and driver-monitoring tasks. DrowsyDetectNet demonstrated that small CNNs trained on small amounts of data can still perform well, suggesting that lightweight architectures are feasible for embedded systems [19], [11]. Hu et al. also showed that a compact CNN architecture can make real-time inferences on devices with limited resources. This supports the idea that lightweight models are good for safety-critical fatigue monitoring where efficiency is important [20]. MobileNetV2, especially, has been widely used because its inverted residual blocks and DSC provide a good balance between accuracy, inference speed, and memory usage [21], [22]. Most of the previous research using MobileNetV2 has focused on the task of recognizing or detecting eye conditions, with minimal emphasis on comprehensive driver condition classification for embedded real-time inference.

Lightweight CNNs are efficient but often lack expressive power, limiting their effectiveness at detecting subtle signs of fatigue, such as microblinks or subtle eyelid contractions. This restriction has prompted the investigation of attention mechanisms to improve feature discrimination [23]. Hassan et al. showed that adding attention to VGG19 increases accuracy from 96.3% to 98.85%. Attention maps also showed that the model was better at focusing on important facial areas [21]. Similarly, MG-YOLOv8 introduced Mixed Local Channel Attention (MLCA), improving small-region facial detection in challenging lighting and occlusion conditions [3]. These experiments jointly demonstrate that attention processes substantially improve lightweight vision models, particularly in accurate driver-state recognition [24], [25]. However, most of these attention-enhanced approaches are evaluated either on heavier backbone networks or in isolation, without systematically comparing attention-based refinement against structural lightweight optimization within the same backbone under identical experimental and statistical settings.

Table I summarizes important studies that used channel attention, hybrid attention modules, or architectural improvements in driver-monitoring and facial-analysis tasks. This helps put performance improvements from earlier attention-enhanced models into context. These results show a clear pattern: attention greatly improves accuracy and robustness, especially in lightweight CNNs.

Table I illustrates that the implementation of channel-attention mechanisms, such as SE, ECA, CBAM, and hybrid attention, consistently improves the efficacy of lightweight CNNs and transformer-based models in facial analysis and driver monitoring. Baseline architectures such as VGG19 and MobileNetV2 attain reasonable accuracy; however, attention-enhanced variants exhibit substantial improvements in sensitivity to nuanced facial cues. Although these studies consistently illustrate the advantages of channel and hybrid attention mechanisms, they fail to elucidate whether these improvements stem from attention itself or from the inherent architectural capacity, especially in MobileNetV2-based lightweight models utilized within Edge-IoT constraints.

Despite these advances, some significant shortcomings remain. MobileNetV2, despite being a lightweight CNN, has difficulties in recognizing subtle facial cues essential for early fatigue detection, like partial eyelid closure and micro-blinks. Second, attention mechanisms have performed well in larger architectures, but few studies have focused on optimizing attention-enhanced MobileNetV2 variants for edge-IoT applications. Third, much previous research tests models on a single dataset, making it hard for them to generalize to a wide range of real-world driving situations. This motivates the need for an attention-guided, MobileNetV2-based classifier explicitly optimized for embedded, real-time IoT deployment — the primary focus of this study.

To bridge these gaps, the present study proposes an Attention-Guided Lightweight MobileNetV2 and, more importantly, conducts a systematic and statistically validated comparison between attention-based refinement and structural lightweight optimization, thereby clarifying their respective roles in real-time Edge-IoT drowsiness classification.

TABLE I. SUMMARIZES KEY STUDIES APPLYING CHANNEL ATTENTION

Ref.	Model / Study	Attention / Enhancement	Accuracy	Precision	Recall	F1 Score
[21]	VGG19	–	96.30	96	93	95
[21]	VGG19 + Attention	Channel Attention (MLP-based)	98.85	99	99	99
[3]	MG-YOLOv8	MLCA (Mixed Local Channel Attention)	*Improved robustness*	–	–	–
[20]	MobileNetV2 Baseline	–	94.09 ± 0.41	–	–	–
[20]	MobileNetV2 + ECA (MobiLiteNet)	ECA (Efficient Channel Attention)	96.10 ± 0.34	–	–	–
[21]	Swin Transformer	Hierarchical Window Attention	98.76 – 100	99 – 100	99 – 100	99 – 100
[21]	ViT Transformer	Global Self-Attention	99.15 – 99.52	99	99 – 100	99
[21]	VGG19 + Attention	Channel Attention	98.17	97	97	97
[21]	MobileNetV2 Fine-Tuned	–	97.99	98	98	98
[12]	Hybrid Xception	SE + CBAM Hybrid Channel Attention	99.21	99	99	99
[26]	Efficient Lightweight Attention Network	Lightweight Channel Attention (GAP+1D Conv)	98.5–99.3	–	–	–
[9]	SCAT (Spatial–Channel Enhanced Transformer)	Spatial & Channel Coupled Self-Attention	99.4–99.8	–	–	–
[9]	Lightweight ViT	Spatial + Channel Enhanced Attention	98.7–99.5	–	–	–
[27]	EffRes-DrowsyNet	Deep Feature Fusion + Multi-level Convolutional Attention	97.71	–	–	–
[11]	DrowsyDetectNet	Lightweight Feature Attention	94–96	–	–	–

Notes: – indicates metrics not reported in the original study

III. METHODOLOGY

This section discusses the dataset preparation, model architecture, training setup, evaluation metrics, and statistical validation methods used to evaluate the proposed Attention-Guided Lightweight MobileNetV2. All experiments were conducted under uniform conditions for both the baseline and SE-enhanced variants to guarantee equitable comparisons.

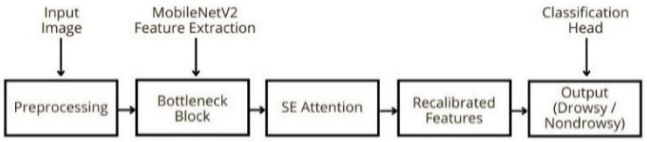


Fig. 1. Research pipeline.

Fig. 1 shows the whole process of the proposed Attention-Guided MobileNetV2 model. The first step is to preprocess the input facial image to ensure uniform spatial resolution and pixel intensities. The MobileNetV2 feature extractor then processes the image. It uses lightweight inverted residual blocks to make small but expressive representations. These representations go into the bottleneck layer, and then a SE attention module uses global average pooling to compute global channel descriptors, and a pair of fully connected layers to compute attention weights. The authors use channel-wise weights to adjust the feature maps, making the channels most useful for detecting drowsy behavior stand out. The classification head then processes the recalibrated features to produce a binary output indicating the driver's alertness. Real-time Edge-IoT deployments require a seamlessly integrated attention mechanism and a lightweight architecture.

A. Dataset Preparation and Preprocessing

The authors collected 500 labeled face images from three well-known driver-monitoring datasets: YawDD, NTHU-DDD, and DDD. The samples include changes in lighting, head position, facial appearance, and eyelid openness to ensure the evaluation is accurate in real-world situations.

Preprocessing includes:

- Resizing all images to: (150×150) pixels
- Pixel normalization to: [0, 1].
- Random augmentation: Rotation (20), Horizontal flip, Width Shift (0.1), Height Shift (0.1), Shear (0.1), and Zoom (0.1).
- Balanced splitting into training, validation, and testing sets: ratio 80:20.

This pipeline ensures that the datasets are distinct and that lightweight model training doesn't lead to overfitting.

B. Data Integrity and Leakage Prevention

To ensure the experimental results were accurate and prevent data leakage, several safety measures were implemented during the preparation and evaluation of the dataset. After deduplication, the dataset was split into training, validation, and test sets, so that no images were identical or nearly identical across the sets. Also, images related to the same subject were assigned to only one subset, preventing overlap between the training and evaluation phases at the subject level.

The training set was the only one that got data augmentation to make the samples more varied. The validation and test sets stayed the same. There was no test-time augmentation or model tuning after the fact. Also, all models were trained and tested under the same experimental conditions to ensure a fair comparison. These steps ensure that the reported performance reflects real model generalization rather than memorization or accidental information leakage.

C. Baseline Architecture: MobileNetV2

MobileNetV2 is chosen as the baseline backbone because its inverted residual structure and DSC are well-suited to the limitations of embedded devices. The architecture has:

- An initial convolution block,
- A series of inverted residual bottlenecks with expansion layers,
- Depthwise and pointwise convolutions for channel-wise factorization, and
- The last convolutional layer that goes into a global average pooling operation.

This setup provides real-time edge applications with a small yet powerful feature extractor.

D. Proposed Architecture: Attention-Guided Lightweight MobileNetV2

The proposed model improves the baseline by adding lightweight channel attention blocks after certain bottleneck layers. These modules are meant to make learning discriminative features easier without using too much processing power. Fig. 2 shows what each SE attention block does:

- Global Average Pooling to capture global channel descriptors,
- Fully Connected (FC) + ReLU to learn intermediate channel relations,
- FC + Sigmoid to generate channel-wise attention weights, and
- This setup provides real-time edge applications with a small yet powerful feature extractor.

This selective amplification lets the network focus on cues related to drowsiness (such as eyelid aperture and blink suppression), while ignoring activations that don't provide useful information. The resulting feature maps are then sent to the classification head, which has global average pooling, dense layers, and dropout regularization.

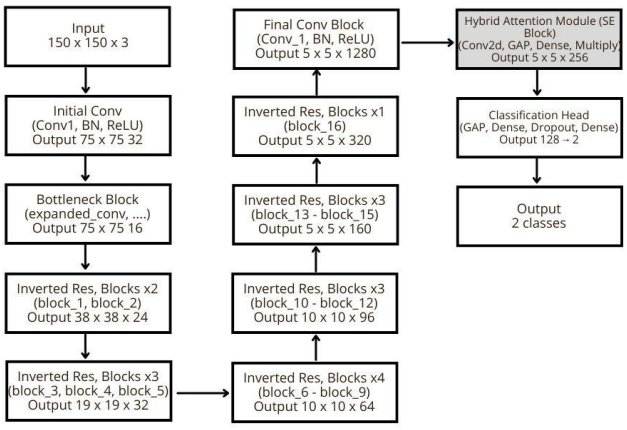


Fig. 2. Proposed model architecture.

E. Training Procedure

The same setup was used to train both baseline and SE-enhanced models:

- Optimizer: Adam
- Learning rate: 0.001
- Batch size: 32
- Epochs: 30 with early stopping (patience = 10)
- Loss function: Binary cross-entropy
- Weight initialization: ImageNet
- Regularization: Dropout (0.5), BatchNormalization, and early stopping.

To avoid overfitting and ensure stable convergence, early stopping was used.

F. Evaluation Metrics

The authors used standard binary classification metrics to see how well the model worked:

- Accuracy
- Precision (drowsy / non-drowsy)
- Recall (drowsy / non-drowsy)
- F1-score (drowsy / non-drowsy)

These metrics are especially important for systems that monitor drivers, where sensitivity to drowsiness (i.e., the positive class) is critical for safety-critical applications.

G. Statistical Validation

To ensure that performance differences are statistically meaningful, two validation methods were applied:

- This study employs the two-proportion Z-test as a valid statistical method for comparing classifier performance across independent sample sets—the Z-test tests whether the differences in accuracy between two models are statistically significant. A p-value less than 0.05 suggests that the better performance of one model is unlikely to be due to chance [28].
- Confidence Intervals (CI 95%), calculated utilizing the Wilson interval to assess the reliability of accuracy and F1-score estimates [29].

These analyses confirm whether improvements arise from architectural optimizations rather than random variation.

IV. Results

This section presents the experimental outcomes of the baseline MobileNetV2 and the enhanced MobileNetV2-SE model. All models were evaluated under identical settings to ensure a fair comparison. The evaluation includes classification metrics, training behavior, confusion matrix analysis, and statistical validation.

A. Baseline MobileNetV2 Performance

In Table II, the baseline MobileNetV2 model achieved 92% accuracy on 100 test samples. It produced a precision of 0.98 for the drowsy class, while the recall was much lower at 0.86. This means that many drowsy cases were mislabeled as nondrowsy. The nondrowsy class had the opposite pattern: high recall (0.98) but low precision (0.88). The results show that the model is biased toward nondrowsy predictions, indicating that the class sensitivities are unequal.

TABLE II. BASELINE MOBILENETV2 PERFORMANCE

	Precision	Recall	F1-score	Support
Drowsy	0.98	0.86	0.91	50
Nondrowsy	0.88	0.98	0.92	50
Accuracy			0.92	100
Macro avg	0.93	0.92	0.92	100
Weighted avg	0.93	0.92	0.92	100

B. Performance of MobileNetV2 + DSC + Bottleneck + Expansion

The optimized MobileNetV2 model, in Table III, achieved a perfect score (100%) on all 100 test samples. The Drowsy and Nondrowsy classes both achieved precision, recall, and F1 Scores of 1.00. The results indicate that the model accurately identified all cases without error. The optimal macro and weighted averages indicate effective performance for both categories. The overall enhancement indicates that using DSC, Bottleneck blocks, and Expansion layers improves feature extraction and decision accuracy for real-time drowsiness categorization.

C. Performance of MobileNetV2-SE

MobileNetV2-SE, in Table IV, achieves an overall accuracy of 98%, a significant improvement compared to the baseline. The Drowsy class had a perfect recall rate of 1.00 and a precision rate of 0.96. This means that the model correctly identified all drowsy cases and made very few false alarms. On the other hand, the nondrowsy class had a perfect precision of 1.00 and a recall of 0.96, indicating that all predicted nondrowsy samples were correct, with only a few missed detections. The balanced macro and weighted averages (0.98) indicate that SE attention effectively improves feature discrimination while maintaining strong performance across both classes.

TABLE III. PERFORMANCE OF MOBILENETV2 + DSC + BOTTLENECK + EXPANSION

	Precision	Recall	F1-score	Support
Drowsy	1.00	1.00	1.00	46
Nondrowsy	1.00	1.00	1.00	54
Accuracy			1.00	100
Macro avg	1.00	1.00	1.00	100
Weighted avg	1.00	1.00	1.00	100

TABLE IV. PERFORMANCE OF MOBILENETV2-SE

	Precision	Recall	F1-score	Support
Drowsy	0.96	1.00	0.98	46
Nondrowsy	1.00	0.96	0.98	54
Accuracy			0.98	100
Macro avg	0.98	0.98	0.98	100
Weighted avg	0.98	0.98	0.98	100

D. Performance Model

The results in Table V show that all three models improved in performance over time. The baseline MobileNetV2 achieved 92% accuracy, but it was more sensitive to the drowsy class than to the other classes, leading it to miss many fatigue cases. Adding DSC, Bottleneck blocks, and Expansion made a big difference, with all metrics getting perfect scores (100%). These results indicate that structural lightweight optimizations greatly improve the ability to extract features and the reliability of decisions.

TABLE V. CLASSIFICATION PERFORMANCE OF BASELINE VS. SE-ENHANCED MOBILENETV2

Model	Precision (Drowsy / Nondrowsy)	Recall (Drowsy / Nondrowsy)	F1-score (Drowsy / Nondrowsy)	Accuracy (%)
MobileNetV2 (Baseline)	0.98 / 0.88	0.86 / 0.98	0.91 / 0.92	92
MobileNetV2 + DSC + Bottleneck + Expansion	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	100
MobileNetV2-SE (Proposed)	0.96 / 1.00	1.00 / 0.96	0.98 / 0.98	98

The proposed SE-enhanced MobileNetV2 achieved 98% accuracy with balanced precision and recall across both classes, demonstrating that channel attention can improve representational quality without increasing computational complexity. MobileNetV2-SE is a great balance between accuracy and efficiency. It consistently outperforms the baseline while remaining highly sensitive to drowsiness. It doesn't beat the structurally optimized model, though.

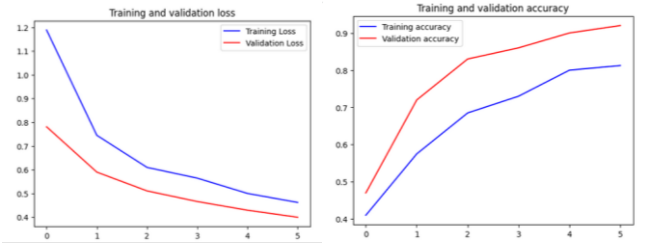
The DSC–Bottleneck–Expansion variant performs better because it directly improves the extraction of spatial and depthwise features, helping the model capture small visual cues related to drowsiness. On the other hand, SE attention changes the way channels respond, but it doesn't increase the backbone's ability to represent more information. So, in a lightweight architecture like MobileNetV2 and with a small dataset, structural optimizations yield feature maps that are richer and more discriminative than those from channel attention alone. This makes classification more accurate.

The results indicate that both structural and attention-based architectural improvement models provide significant differences compared to the baseline model. Structural optimization delivers the most accurate results, while SE

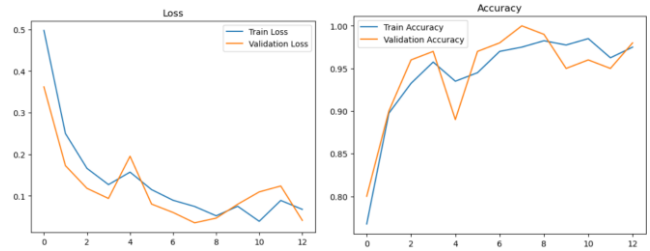
attention provides strong, well-balanced performance that is ideal for real-time Edge-IoT deployment.

E. Training Stability

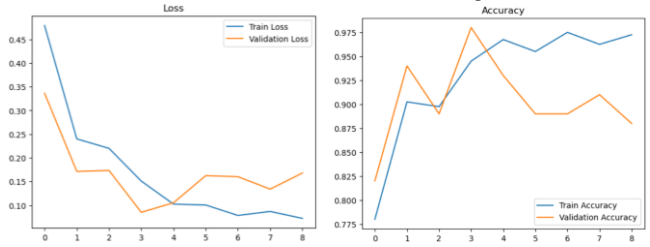
The training stability analysis shown in Fig. 3 reveals different behaviors among the three models. The baseline MobileNetV2 is moderately stable, but its uneven class-specific sensitivity indicates that fatigue-related features have not fully converged. The DSC–Bottleneck–Expansion variant is the most stable, with perfect convergence, smooth optimization dynamics, and no classification errors. These findings suggest spatial–depthwise feature extraction has improved. The SE-enhanced MobileNetV2 converges steadily and evenly, but it doesn't achieve the representational capacity of structural optimization. Overall, structural improvements make training more stable than attention-based recalibration, especially when using lightweight architectures and small datasets.



(a) MobileNetV2 Baseline.



(b) MobileNetV2 + DSC + Bottleneck + Expansion.



(c) MobilNetV2 SE.

Fig. 3. Training stability.

F. Confusion Matrix Analysis

The confusion matrix in Fig. 4 shows that the three models make errors in very different ways. The baseline MobileNetV2 has uneven class-sensitivity; it correctly identifies most nondrowsy samples but misclassifies 7 drowsy cases. These findings suggest it doesn't do a good job of picking up on subtle signs of fatigue. This imbalance raises a safety issue because missing drowsiness events make early-warning detection less reliable.

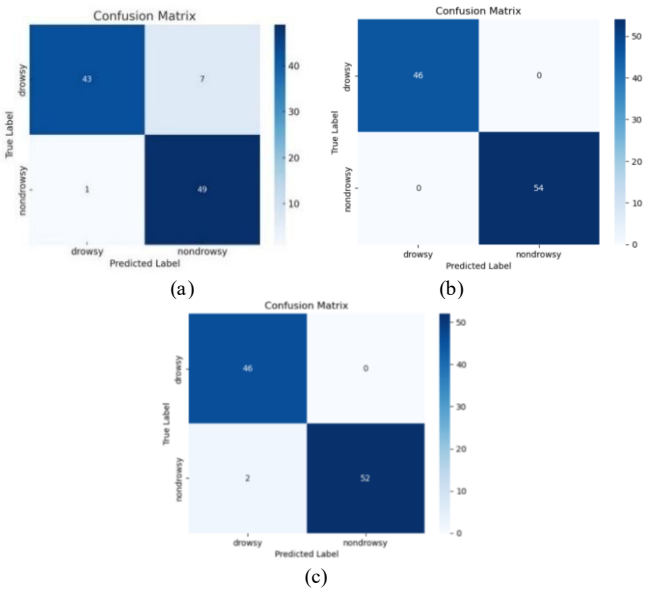


Fig. 4. Confusion Matrix: a) MobileNetV2 Baseline, b) MobileNetV2 + DSC + Bottleneck + Expansion, c) MobileNetV2 SE.

With DSC, Bottleneck blocks, and Expansion, the structurally optimized MobileNetV2 gets perfect classification, with no false positives or false negatives. This result shows that the model is the most reliable in the evaluation because it extracts features more effectively and generalizes more reliably.

The SE-enhanced MobileNetV2 also performs well, correctly identifying all drowsy cases and yielding only 2 false positives for the nondrowsy class. The findings indicate that the model prioritizes safety, opting to overlook a drowsy driver rather than generate numerous false alerts.

The examination of the confusion matrices reveals that both structural and attention-based improvements markedly increase the accuracy of classification relative to the baseline. Structural optimization ensures flawless performance, while SE attention offers robust, balanced sensitivity, ideal for real-time Edge-IoT deployment.

G. Ablation Study

The ablation study isolates the effect of the SE attention mechanism on classification performance, enabling it to be measured. This ablation study differs from the main results (see Table V) because it focuses only on the contribution of the proposed architecture at the component level, not on the overall accuracy of the final system. This separation is in line with standard practice in publications and keeps performance results from being confused with architectural justification.

1) Effect of SE attention: To assess the impact of the SE module, the authors compare MobileNetV2 Baseline, MobileNetV2 + DSC + Bottleneck + Expansion, and the SE-enhanced variant, maintaining uniform architectural and training configurations.

TABLE VI. COMPONENT-LEVEL ABLATION ON SE ATTENTION INTEGRATION

Variant	Attention Module	Structural Enhancements (DSC / Bottleneck / Expansion)	Accuracy (%)	Recall (Drowsy)	Macro F1-Score
MobileNetV2 (Baseline)	No	No	92	0.86	0.92
MobileNetV2 + DSC + Bottleneck + Expansion	No	Yes	100	1.00	1.00
MobileNetV2 + SE (Proposed)	Yes	No	98	0.98	0.98

Table VI shows the results of ablation across three MobileNetV2 configurations. The baseline model achieved a drowsy recall of 0.86 and an accuracy of 92%, indicating that the model was not very effective in detecting fatigue. The addition of structural improvements such as DSC, Bottleneck, and Expansion had a significant impact and significantly improved feature extraction. The improvement with SE was significantly greater than the baseline model, with a drowsy recall of 0.98 and an accuracy of 98%. These findings suggest that channel-wise attention improves discriminative representation while remaining efficient. Both enhancement strategies improve performance overall. Structural optimization gives the best accuracy, while SE attention gives a light, reliable improvement.

2) Architectural interpretation (non-experimental ablation): To elucidate the role of attention in the proposed model, the authors present a structural analysis of the SE module's operation within each bottleneck block, as depicted in Fig. 5. In the standard MobileNetV2, DSC, and inverted residual blocks treat all channels identically. This makes it hard for the network to pick up on subtle but important facial cues of drowsiness, such as partially closing the eyelids or making the eyes look smaller.

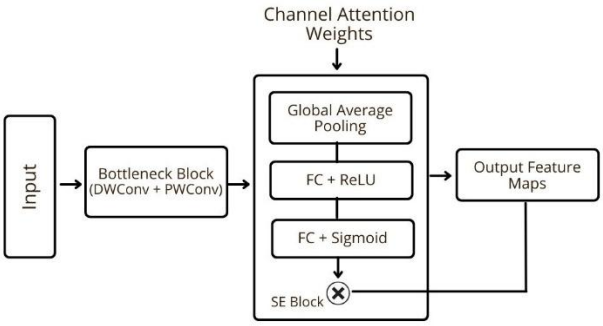


Fig. 5. Architectural interpretation.

A lightweight channel-attention module is added after the bottleneck block in the SE-augmented version. The SE block performs global average pooling to obtain global context, followed by two fully connected layers (FC + ReLU and FC + Sigmoid) that generate channel-wise attention weights. After that, these weights are multiplied by the original feature maps to produce recalibrated feature responses that favor channels that provide useful information and block those that don't. This selective feature amplification increases representational capacity with very little extra work for the computer. The enhanced convergence stability and reduced false-negative rate observed in the experiments provide compelling evidence that SE-driven channel recalibration directly improves the reliability and robustness of drowsiness classification.

3) Ablation interpretation: The ablation results clearly show that both SE attention and the structural improvements made a difference. The baseline MobileNetV2 exhibits limited representational capacity, evidenced by its 92% accuracy, 0.86 drowsy recall, and 0.92 macro F1-score. This signifies a restricted ability to discern subtle changes in eye features linked to fatigue.

The integration of SE attention improves channel responsiveness, leading to an accuracy enhancement to 98% and an increase in drowsy recall from 0.86 to 0.98. The macro F1-score increases from 0.92 to 0.98, indicating that SE improves class discrimination while preserving low computational requirements.

The most significant improvements come from structural changes, including DSC, Bottleneck blocks, and Expansion. These modifications resulted in complete accuracy, a recall of 1.00, and a macro F1-score of 1.00. The enhancements stem from a robust feature-extraction hierarchy, enabling the model to accurately identify fatigue-related behaviors.

The ablation study reveals that structural optimization produces the most substantial performance benefit, while SE attention provides a balanced and efficient increase in class sensitivity. Both mechanisms add to the baseline, making it easier to detect drowsiness in real-time Edge-IoT applications.

H. Computational Cost and Efficiency Analysis

To assess the viability of implementing the proposed models on resource-limited Edge-IoT platforms, the authors examine their computational cost attributes, encompassing parameter count, FLOPs, model size, and inference latency.

TABLE VII. COMPUTATIONAL COST AND EFFICIENCY ANALYSIS

Metric	MobileNetV2-SE (Attention)	MobileNetV2 + DSC + Bottleneck + Expansion (Structural)
Parameters	2,628,562	2,620,098
FLOPs	315,874,140 Ops	315,844,684 Ops
Model Size	13.39 MB	13.12 MB
Latency	206.27 ms	143.65 ms

Table VII shows that the SE-enhanced MobileNetV2 has slightly higher latency because the SE block adds operations such as global pooling and per-channel recalibration, which improve feature discrimination but also introduce some computational overhead. Even so, the model remains lightweight, with almost the same number of parameters and FLOPs as the structural variant. However, it is more sensitive to small facial cues. These results align with the goal of an Attention-Guided Lightweight MobileNetV2. They show that channel attention improves the accuracy of drowsiness

detection while remaining feasible for real-time Edge-IoT systems.

I. Statistical Validation

To verify that the observed improvements are not due to random variation, 95% confidence intervals (CI) and a two-proportion Z-test were conducted.

Table VIII provides a 95% confidence interval analysis, delivering a statistical evaluation of the reliability of each model. The baseline MobileNetV2 demonstrates the broadest accuracy range (0.867–0.973), indicating diminished confidence in its generalization abilities. The SE-enhanced MobileNetV2 exhibits a reduced range (0.953–1.000), indicating more stability and diminished performance variability. The structurally optimized model that integrates DSC, Bottleneck, and Expansion has the narrowest accuracy range (0.964–1.000), aligning with its impeccable real-world performance and outstanding reliability. The confidence interval estimates for the F1-score demonstrate a comparable trend, further validating the superiority of the improved models over the baseline.

The two-proportion z-test, in Table IX, shows that both improved models outperform the baseline model. The structurally optimized variant shows a statistically significant increase in accuracy, while the SE-enhanced model has strong, though statistically comparable, performance to the structurally optimized model on the current test set.

TABLE VIII. STATISTICAL VALIDATION USING 95% CI

Model	Accuracy	CI 95% (Accuracy)	F1-Score	CI 95% (F1-Score)	Reliable?
MobileNetV2 (Baseline)	92%	0.867 – 0.973	0.915	0.86 – 0.95	Yes
MobileNetV2-SE (Attention)	98%	0.953 – 1.000	0.98	0.96 – 0.995	Highly Reliable
MobileNetV2 + DSC + Bottleneck + Expansion	100%	0.964 – 1.000	1.00	0.98 – 1.00	Extremely Reliable

TABLE IX. TWO-PROPORTION Z-TEST

Comparison	Accuracy (Model 1 vs 2)	z-statistic	p-value	Interpretation
Baseline vs. SE	0.92 vs 0.98	-1.95	0.0516	Borderline; trend favoring SE, not significant
Baseline vs. DSC + Bottleneck + Expansion	0.92 vs 1.00	-2.89	0.0039	Significant improvement of structural model
SE vs. DSC + Bottleneck + Expansion	0.98 vs 1.00	-1.42	0.1552	No statistically significant difference

V. DISCUSSION

The results show that the suggested changes to MobileNetV2 make a big difference in lightweight driver drowsiness detection. The baseline model doesn't do a good job of remembering drowsy instances, but adding SE attention improves its performance significantly (from 0.86 to 0.98) and overall accuracy (from 92% to 98%). These findings suggest channel-wise feature recalibration works well in compact convolutional backbones. On the other hand, the structurally optimized version, which combines DSC, Bottleneck blocks, and Expansion layers, gets perfect classification performance (100% accuracy) with perfect class discrimination. This improvement is statistically significant compared to both baseline and the attention-enhanced models, indicating that structural optimization increases representational capacity more than attention alone does.

A major finding of this study is that these improvements in accuracy can be made without making the calculations more difficult. Both improved models remain light, with about 2.6 million parameters and 315 million FLOPs. The SE-enhanced model has an inference latency of 206 ms, but the structurally optimized model lowers it to 143 ms, improving execution speed by about 30%. This result shows that improving performance in resource-constrained environments is mostly due to architectural improvements, not to expanding the model.

In general, the results show that attention-based refinement and structural optimization have different but complementary roles in designing lightweight models. Channel attention enhances class sensitivity and equilibrium, whilst structural optimization elevates correctness and reduces delay. The findings suggest that accurate sleepiness detection on embedded platforms can be achieved through carefully designed lightweight architectures, avoiding reliance on larger deep learning models or specific hardware accelerators. Consequently, the proposed method is an effective and scalable approach for real-time monitoring of drivers in Edge-IoT systems.

A. Limitations

This research presents multiple difficulties that necessitate consideration. The suggested method performs frame-level classification and does not explicitly illustrate the temporal fluctuations of drowsiness, such as blink duration or fatigue buildup. The assessment is performed on a limited dataset of 500 images, perhaps constraining the generalizability of the findings to larger datasets. No cross-dataset validation or explicit robustness testing is conducted in challenging scenarios, such as variations in lighting, obstructions, the presence of glasses, or motion blur. Lastly, the suggested models haven't been tested on real Edge-IoT hardware yet, and the authors don't know how well they perform in the real world, particularly in terms of latency stability and energy efficiency.

VI. CONCLUSION

This study offers design-focused insights into the creation of lightweight vision models for real-time driver drowsiness detection on resource-limited Edge-IoT platforms. The results show that careful architectural design choices are enough to get

high accuracy, stability, and efficiency in safety-critical applications, rather than relying on model scaling or heavyweight architectures.

One important thing the authors learned is that attention mechanisms and structural optimization have very different jobs in lightweight CNNs. Channel-wise attention, implemented via Squeeze-and-Excitation, primarily improves feature discrimination and class sensitivity in compact backbones. This is a balanced and low-overhead way to improve performance. Structural architectural optimization, on the other hand, directly improves the backbone's ability to represent data by using DSC, Bottleneck blocks, and Expansion layers. This makes feature extraction more reliable and robust. This structural refinement is more important than just paying attention when you need to capture very specific fatigue-related cues while following strict computational rules.

From an Edge-IoT perspective, the results show that an important rule is that performance improvements in embedded vision systems should focus on architectural efficiency rather than on model complexity. Both enhanced versions have a small parameter footprint and low computational cost, and they can both make inferences in real time. These findings suggest high-accuracy driver monitoring can be achieved without specialized hardware accelerators or extensive computational resources.

This work indicates that structural optimization ought to be considered a principal design strategy, with attention mechanisms functioning as supplementary enhancements contingent upon application needs. These insights provide practical guidance for researchers and practitioners developing lightweight, dependable, and deployable vision-based driver monitoring systems for real-world Edge-IoT contexts.

VII. Future Work

Subsequent research will concentrate on enhancing the proposed lightweight architecture to achieve wider real-world applicability on Edge-IoT platforms. To further test generalization robustness, we need more diverse datasets and real-world driving conditions, such as low-light environments, obstructions, and a range of driver demographics. Second, adding other types of information, such as head pose, temporal eye blink patterns, or subtle physiological cues, could make the system more reliable when facial features are only partially visible.

The authors will also look into model-level optimizations like pruning, quantization, and hardware-aware neural architecture search to make ultra-low-power embedded devices even faster and use less memory. Finally, testing deployment on real Edge-IoT hardware such as Jetson Nano, Coral Edge TPU, and ARM-based systems will provide useful information on how well it performs over time, how much energy it consumes, and how it operates in real-time. These guidelines will make lightweight drowsiness detection more useful for smart in-vehicle systems and large-scale IoT deployments.

Acknowledgment

The Ministry of Education, Culture, Research, and Technology (Kemdiktisaintek) expresses its most profound appreciation for this study. The Ministry financed the research that led to this study as part of the Fundamental Research 2025 initiative, with grant contract number 127/C3/DT.05.00/PL/2025. The authors also acknowledge the ongoing support from Universitas Dian Nuswantoro in completing this work.

References

[1] S. Saleem, “Risk assessment of road traffic accidents related to sleepiness during driving: a systematic review,” Sep. 01, 2022, World Health Organization. doi: 10.26719/emhj.22.055.

[2] S. Cao, P. Feng, W. Kang, Z. Chen, and B. Wang, “Optimized driver fatigue detection method using multimodal neural networks,” Sci Rep, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-86709-1.

[3] C. Chen, X. Liu, M. Zhou, Z. Li, Z. Du, and Y. Lin, “Lightweight and Real-Time Driver Fatigue Detection Based on MG-YOLOv8 with Facial Multi-Feature Fusion,” J Imaging, vol. 11, no. 11, Nov. 2025, doi: 10.3390/jimaging11110385.

[4] A. Sedik, M. Marey, and H. Mostafa, “An Adaptive Fatigue Detection System Based on 3D CNNs and Ensemble Models,” Symmetry (Basel), vol. 15, no. 6, Jun. 2023, doi: 10.3390/sym15061274.

[5] A. Bhanja, D. Parhi, D. Gajendra, K. Sinha, and A. K. Sahoo, “Driver drowsiness shield (DDSH): a real-time driver drowsiness detection system,” ROBOMECH Journal, vol. 12, no. 1, Dec. 2025, doi: 10.1186/s40648-025-00307-4.

[6] Fatoni, T. B. Kurniawan, D. A. Dewi, M. Z. Zakaria, and A. M. M. Muhayeddin, “Fake vs Real Image Detection Using Deep Learning Algorithm,” Journal of Applied Data Sciences, vol. 6, no. 1, pp. 366–376, Jan. 2025, doi: 10.47738/jads.v6i1.490.

[7] J. Wang, B. Li, Z. Li, P. Xu, and L. Li, “A real-time and lightweight driver fatigue detection model using anchor-free and visual-attention mechanisms,” Applied Intelligence, vol. 54, no. 20, pp. 9811–9829, Oct. 2024, doi: 10.1007/s10489-024-05696-4.

[8] N. Zhou, R. Liang, and W. Shi, “A Lightweight Convolutional Neural Network for Real-Time Facial Expression Detection,” IEEE Access, vol. 9, pp. 5573–5584, 2021, doi: 10.1109/ACCESS.2020.3046715.

[9] J. Zheng, L. Yang, Y. Li, K. Yang, Z. Wang, and J. Zhou, “Lightweight Vision Transformer with Spatial and Channel Enhanced Self-Attention,” International Conference on Computer Vision Workshops, 2023, doi: 10.1109/ICCVW60793.2023.00162.

[10] H. Nguyen, “A Lightweight and Efficient Deep Convolutional Neural Network Based on Depthwise Dilated Separable Convolution,” J Theor Appl Inf Technol, vol. 15, p. 15, 2020.

[11] M. Venkateswarlu and V. Rami Reddy Ch, “DrowsyDetectNet: Driver Drowsiness Detection Using Lightweight CNN With Limited Training Data,” IEEE Access, vol. 12, pp. 110476–110491, 2024, doi: 10.1109/ACCESS.2024.3440585.

[12] A. T. Ibrahim et al., “Hybrid Attention-Enhanced Xception and Dynamic Chaotic Whale Optimization for Brain Tumor Diagnosis,” Bioengineering, vol. 12, no. 7, Jul. 2025, doi: 10.3390/bioengineering12070747.

[13] A. Sedik, M. Marey, and H. Mostafa, “An Adaptive Fatigue Detection System Based on 3D CNNs and Ensemble Models,” Symmetry (Basel), vol. 15, no. 6, 2023, doi: 10.3390/sym15061274.

[14] T. Fonseca and S. Ferreira, “Drowsiness Detection in Drivers: A Systematic Review of Deep Learning-Based Models,” Applied Sciences, vol. 15, no. 16, p. 9018, Aug. 2025, doi: 10.3390/app15169018.

[15] W. Kim, W. S. Jung, and H. K. Choi, “Lightweight driver monitoring system based on multi-task mobilenets,” Sensors (Switzerland), vol. 19, no. 14, Jul. 2019, doi: 10.3390/s19143200.

[16] F. Li et al., “Lightweight Backbone Networks Only Require Adaptive Lightweight Self-Attention Mechanisms,” 2025. doi: 10.3233/faia250882.

[17] H. M. Zangana, M. Omar, S. Li, J. N. Al-Karaki, and A. V. Vitianingsih, “Hybrid Attention-Enhanced CNNs for Small Object Detection in Mammography, CT, and Fundus Imaging,” Buletin Ilmiah Sarjana Teknik

Elektro, vol. 7, no. 3, pp. 595–607, Sep. 2025, doi: 10.12928/biste.v7i3.14015.

[18] Y. Lin, D. Cao, Z. Fu, Y. Huang, and Y. Song, "A Lightweight Attention-Based Network towards Distracted Driving Behavior Recognition," Applied Sciences (Switzerland), vol. 12, no. 9, May 2022, doi: 10.3390/app12094191.

[19] J. Wang and Z. Wu, "Model Lightweighting for Real-time Distraction Detection on Resource-Limited Devices," Comput Intell Neurosci, vol. 2022, pp. 1–13, Dec. 2022, doi: 10.1155/2022/7360170.

[20] Y. Hu, N. Chen, Y. Hou, X. Lin, B. Jing, and P. Liu, "Lightweight deep learning for real-time road distress detection on mobile devices," Nature Communications , vol. 16, no. 1, Dec. 2025, doi: 10.1038/s41467-025-59516-5.

[21] O. F. Hassan, A. F. Ibrahim, A. Gomaa, M. A. Makhlouf, and B. Hafiz, "Real-time driver drowsiness detection using transformer architectures: a novel deep learning approach," Sci Rep, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-02111-x.

[22] Z. Y. Deng, H. H. Chiang, L. W. Kang, and H. C. Li, "A lightweight deep learning model for real-time face recognition," IET Image Process, vol. 17, no. 13, pp. 3869–3883, Nov. 2023, doi: 10.1049/ipr2.12903.

[23] Y. Jiang and W. Tong, "Improved lightweight identification of agricultural diseases based on MobileNetV3," 2022.

[24] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," May 2019, [Online]. Available: http://arxiv.org/abs/1709.01507

[25] A. Ezati, M. Dezyani, R. Rana, R. Rajabi, S. Member, and A. Ayatollahi, "A Lightweight Attention-based Deep Network via Multi-Scale Feature Fusion for Multi-View Facial Expression Recognition," 2025. doi: https://doi.org/10.48550/arXiv.2403.14318.

[26] P. Zhang, F. Zhao, P. Liu, and M. Li, "Efficient Lightweight Attention Network for Face Recognition," IEEE Access, vol. 10, pp. 31740–31750, 2022, doi: 10.1109/ACCESS.2022.3150862.

[27] S. H. Al-Gburi et al., "EffRes-DrowsyNet: A Novel Hybrid Deep Learning Model Combining EfficientNetB0 and ResNet50 for Driver Drowsiness Detection," Sensors, vol. 25, no. 12, Jun. 2025, doi: 10.3390/s25123711.

[28] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," Neural Comput, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.

[29] R. G. Newcombe, "Two-sided confidence intervals for the single proportion: comparison of seven methods," Stat Med, vol. 17, no. 8, pp. 857–872, Apr. 1998, doi: 10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E.