The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 5, No. 1. Online ISSN: 2831-4794

CycleGAN-Based Drunk Synthesis and Attention-Enhanced MobileNetV2 for Driver State Recognition

Linh T.P. Le[1*]
[1]School of Computer Science and Engineering, Vietnam National University Ho Chi Minh City, International University, Ho Chi Minh City, Vietnam
*Corresponding author. E-mail: hyylinh2912@gmail.com

ABSTRACT

Driver monitoring plays a critical role in intelligent transportation systems, yet detecting fatigue, alcohol impairment, and distraction remains highly challenging due to their subtle and overlapping cues. A major obstacle is the scarcity of annotated alcohol-related data, which limits the training of robust classifiers. This work presents a practical framework that combines adversarial data augmentation with lightweight deep learning for unified driver state recognition. A CycleGAN-based pipeline synthesizes alcohol-impaired facial images from fatigue samples, introducing physiologically motivated effects such as skin flushing, gaze irregularities, and periocular redness. For classification, a MobileNetV2 backbone enhanced with Squeeze-and-Excitation (SE) attention adaptively emphasizes critical channels while remaining computationally efficient. Evaluated on a curated seven-class dataset, the model achieves 97.67% accuracy with a test loss of 0.0655, confirming the viability of adversarial augmentation coupled with lightweight CNNs for real-time driver monitoring.

Keywords: Driver State Monitoring, Synthetic Drunk Data Generation, CycleGAN Augmentation, MobileNetV2, Squeeze-and-Excitation

1. INTRODUCTION

Road safety continues to be a global concern, with driver impairment identified as a leading cause of accidents. Fatigue, alcohol intoxication, and distraction represent distinct but interrelated conditions that severely compromise driving performance. Fatigue impairs vigilance and reaction time, alcohol intoxication alters facial physiology and gaze stability, while distraction diverts attention to secondary tasks such as mobile phone usage or interactions with passengers. Monitoring these states accurately and reliably is therefore essential for developing next-generation intelligent transportation systems.

Despite significant progress in computer vision and machine learning, several limitations remain. First, most prior studies focus on a single impairment type, such as fatigue detection or distraction recognition, without addressing the broader spectrum of risks. This narrow scope reduces the robustness of such systems in complex driving environments. Second, the scarcity of real-world datasets for alcohol-impaired driving presents a fundamental bottleneck. Unlike fatigue or distraction, where annotated datasets are available, alcohol-related data are difficult to obtain due to ethical and logistical constraints. Attempts to overcome this limitation using generative adversarial networks (GANs) have often produced unrealistic samples with limited variability. Finally, subtle differences between impaired and non-impaired states—such as micro-expressions, skin tone variations, or gaze shifts—require models capable of emphasizing fine-grained cues rather than relying solely on global appearance.

To address these challenges, this study introduces a unified framework that combines generative augmentation with discriminative feature learning. A CycleGAN-based pipeline generates realistic alcohol-impaired facial states from non-intoxicated data, enhanced through landmark-guided effects such as skin flushing and

gaze jitter. The generated samples are integrated with curated datasets representing fatigue, distraction behaviors, and safe driving, thus establishing a comprehensive multi-class corpus. For recognition, a lightweight MobileNetV2 backbone augmented with Squeeze-and-Excitation attention adaptively recalibrates channel-wise features to highlight subtle but discriminative cues. This design ensures efficiency suitable for real-time deployment while maintaining high classification accuracy.

The contributions of this work are threefold. (i) We introduce a CycleGAN-based augmentation pipeline specifically designed to synthesize alcohol-impaired facial states from fatigue-domain images, thereby directly tackling the scarcity of real-world intoxication data. (ii) We construct a balanced seven-class dataset integrating safe driving, fatigue, drunk, and distraction behaviors, enriched with synthetic drunk faces generated through adversarial training. (iii) We develop and evaluate a lightweight MobileNetV2 classifier enhanced with Squeeze-and-Excitation blocks, showing that a compact CNN with channel attention, when coupled with adversarial augmentation, achieves 97.67% accuracy and real-time feasibility for driver monitoring.

2. RELATED WORK

Driver monitoring has attracted extensive research across multiple methodological directions, ranging from handcrafted feature engineering to deep neural architectures and generative augmentation techniques. This section reviews representative approaches that have shaped the field and highlights their limitations.

2.1 Traditional Vision-Based Methods

Early driver state recognition relied heavily on handcrafted features derived from facial geometry and eye dynamics. Measures such as the eye aspect ratio (EAR) for blink detection, mouth opening ratio for yawning, and head pose estimation were frequently employed as indicators of drowsiness or distraction. These descriptors were typically classified using Support Vector Machines (SVMs), Hidden Markov Models (HMMs), or Random Forests. Classical methods achieved promising results in controlled environments [1, 2], but they were highly sensitive to illumination changes, camera angle, and subject-specific differences, which hindered scalability to real-world driving scenarios.

2.2 CNN-Based Fatigue and Distraction Recognition

The introduction of convolutional neural networks (CNNs) significantly advanced driver monitoring by enabling automated feature extraction from raw images. Architectures such as VGGNet [3], ResNet [4], and MobileNet [5] have been applied to fatigue recognition, showing substantial improvements in accuracy and robustness compared to handcrafted approaches. Multi-class distraction recognition has also been addressed with CNN-based pipelines on large-scale datasets such as StateFarm and Vicomtech DMD [6, 7]. Despite these gains, CNN models often emphasize global appearance rather than subtle cues such as micro-expressions or partial occlusions, which are critical for differentiating between impaired and safe driving. Furthermore, most CNN-based studies remain task-specific, focusing exclusively on fatigue or distraction instead of developing unified multi-impairment frameworks.

2.3 Generative Augmentation with GANs

A major limitation in alcohol-impaired driver monitoring is the scarcity of annotated datasets, as data collection under intoxicated conditions poses ethical and safety challenges. Generative Adversarial Networks (GANs) have been explored to mitigate this bottleneck by producing synthetic samples. CycleGAN, proposed by Zhu et al. [8], enables unpaired image-to-image translation and has been widely adopted for style transfer and domain adaptation. Related works in facial analysis, such as StarGAN by Choi et al. [9] and conditional GANs by

Mirza et al. [10], demonstrate how generative models can simulate attributes including age, expression, or gaze. However, naïve GAN implementations for driver monitoring often generate artifacts, low-diversity samples, or unstable results. To address this, researchers have incorporated perceptual losses [11] and identity-preserving constraints [12], as well as domain-specific priors like gaze shifts or skin redness, though systematic evaluation in driver monitoring remains limited.

2.4 Attention and Lightweight Architectures

Attention mechanisms have emerged as powerful tools for emphasizing discriminative features in visual recognition. The Squeeze-and-Excitation (SE) block, proposed by Hu et al. [13], adaptively recalibrates channel responses, allowing CNNs to focus on subtle cues that may otherwise be suppressed. Convolutional Block Attention Module (CBAM), introduced by Woo et al. [14], integrates both spatial and channel attention and has been applied in behavior recognition tasks. At the same time, lightweight CNN architectures such as MobileNetV2 by Sandler et al. [15] and ShuffleNet by Zhang et al. [16] have been widely adopted for real-time applications, including in-vehicle systems with limited computational resources. These developments reflect the increasing demand for models that balance efficiency and accuracy in driver monitoring scenarios.

3. METHODOLOGY

The proposed framework follows an end-to-end pipeline that integrates dataset curation, data preprocessing, generative augmentation, and classification with an attention-enhanced lightweight CNN. An overview of the pipeline is shown in Figure. 1.

3.1 Dataset Construction

Multiple public datasets were combined to build a unified multi-class corpus:

- **Fatigue:** Images from the Kaggle Drowsiness Prediction Dataset were employed to capture facial cues of tired drivers.
- **Distraction:** The Vicomtech Driver Monitoring Dataset (DMD) was used to extract four behaviors: phone call, drinking, hair and makeup, and talking to passenger. Thirty frames were uniformly sampled per video to capture different phases of each action.
- **Drunk:** Alcohol-impaired facial images were obtained from two complementary sources. First, a small collection of real intoxicated driver faces was gathered from publicly available repositories, including open-source research collections and online awareness campaigns where facial data were explicitly released for academic use. These samples, while limited in number, provide authentic visual cues such as facial flushing, drooping eyelids, and gaze instability. Second, to address the inherent scarcity of such data, synthetic drunk faces were generated from fatigue-domain images using CycleGAN. The generative process introduced alcohol-related priors—including skin redness, gaze jitter, and red-eye effects—to ensure visual realism while preserving identity information. The combination of authentic and synthetic data yielded a more balanced and diverse representation of the drunk-driving category.
- **Safe Drive:** Neutral driving states were collected from non-impaired subsets of existing datasets. To avoid class imbalance, only 50% of samples from the active driving subset were retained.

This yields seven categories: safe drive, fatigue, drunk, drinking, hair and makeup, phone call, and talking to passenger.

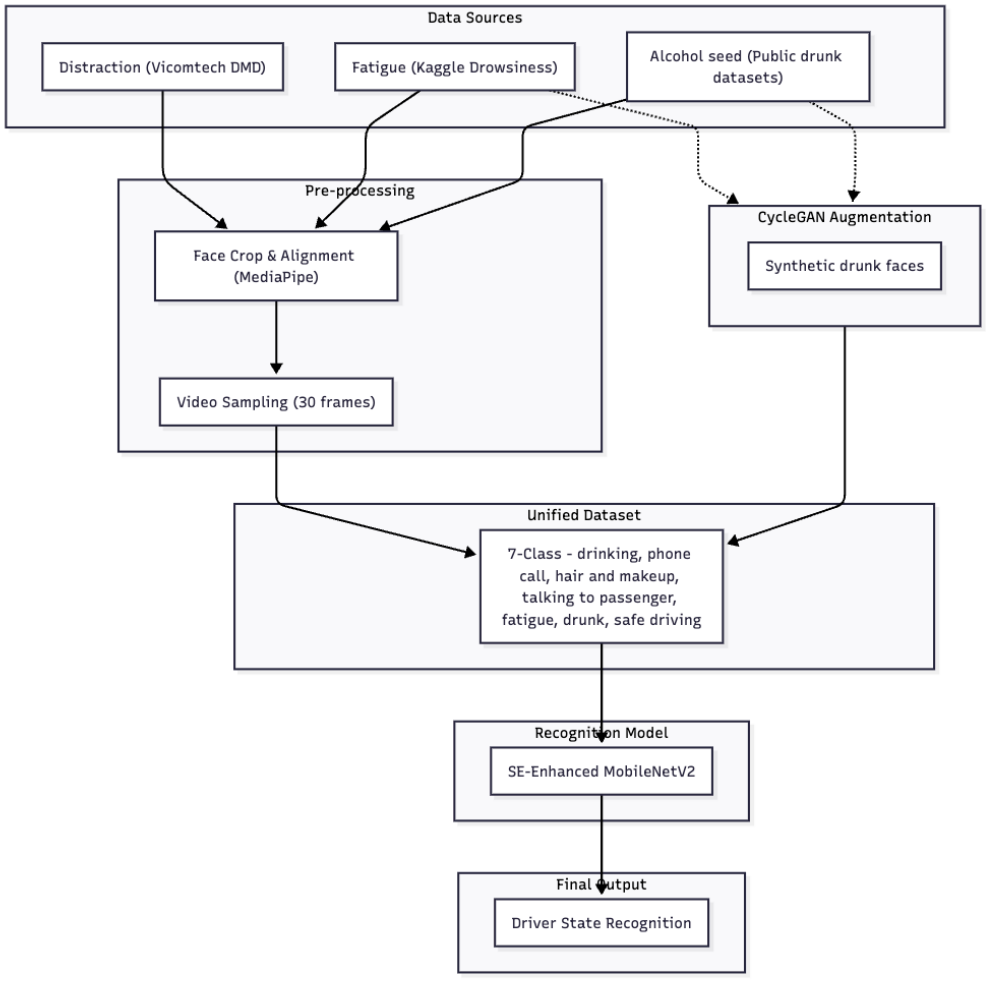


Figure 1: End-to-End pipeline for unified driver state recognition.

3.2 Data Preprocessing

Prior to training, all raw data were passed through a standardized preprocessing pipeline to ensure consistency and robustness across heterogeneous sources. The process consisted of three main stages:

Face detection and cropping. Facial regions were localized using MediaPipe Face Detection, chosen for its efficiency and robustness in real-time scenarios. Detected bounding boxes were used to crop the driver's face, discarding backgrounds that may introduce noise. Frames with failed detections, extreme occlusions, or invalid crops were excluded. For the Vicomtech DMD videos, thirty uniformly spaced frames were sampled from each clip to capture the full temporal variation of distraction behaviors (e.g., lifting a bottle, drinking, lowering).

Data augmentation. To improve generalization and reduce overfitting, augmentation was applied during training. Operations included random resized cropping (scale 0.8–1.0) to simulate varying camera zooms, random horizontal flipping to mimic mirrored perspectives, random rotations up to $\pm 20^{\circ}$ for head pose variation, Gaussian blur ($p = 0.3$) to simulate motion blur or poor camera quality, and color jitter (brightness $\pm 20\%$, contrast $\pm 30\%$) to emulate different lighting conditions. Validation and test sets were resized without augmentation to provide unbiased evaluation. This pre-processing pipeline ensured that the final dataset not only achieved

consistent spatial alignment but also incorporated realistic appearance variations, making the classifier more resilient to real-world driving conditions.

Data splitting strategy. The unified dataset was divided into training, validation, and test sets using a stratified split to preserve class balance. Specifically, 70% of the video sequences were used for training, 15% for validation, and 15% for testing. The split was performed at the sequence level, ensuring that frames extracted from the same video clip were assigned exclusively to a single subset, thereby preventing information leakage. All reported frame-level classification results were obtained on the held-out test set.

3.3 Generative Augmentation with CycleGAN

A fundamental bottleneck in alcohol-impaired driver monitoring lies in the absence of large, annotated datasets, since data collection under intoxicated conditions raises significant ethical and safety issues. To overcome this, this study leverages CycleGAN [8,17] to synthesize drunk-domain faces from fatigue-domain counterparts in an unpaired setting.

The generator is based on ResNet-9 residual blocks, which provide a balance between receptive field size and computational efficiency, while the discriminator adopts a PatchGAN architecture that enforces realism at the patch level. This design allows fine-scale textures, such as skin redness or periocular artifacts, to be preserved more faithfully.

The training objective integrates multiple complementary losses. The adversarial loss encourages realistic synthesis as shown in (1). Cycle-consistency loss, given in (2), stabilizes the bidirectional mapping. Identity loss (3) constrains generators to leave non-target samples unaffected. Perceptual loss (4), based on VGG16 activations [11], enforces high-level semantic similarity. The full objective in (5) balances these terms.

$$\begin{aligned}\mathcal{L}_{\text{GAN}} = \mathbb{E}_{b\sim p_B}[(D_B(b)-1)^2] \\ + \mathbb{E}_{a\sim p_A}[D_B(G_{A\to B}(a))^2].\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L}_{\text{cyc}} = \mathbb{E}_a[\|G_{B\to A}(G_{A\to B}(a)) - a\|_1] \\ + \mathbb{E}_b[\|G_{A\to B}(G_{B\to A}(b)) - b\|_1].\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{\text{id}} = \mathbb{E}_a[\|G_{B\to A}(a) - a\|_1] \\ + \mathbb{E}_b[\|G_{A\to B}(b) - b\|_1].\end{aligned} \tag{3}$$

$$\begin{aligned}\mathcal{L}_{\text{perc}} = \|\phi(G_{A\to B}(a)) - \phi(b)\|_1 \\ + \|\phi(G_{B\to A}(b)) - \phi(a)\|_1.\end{aligned} \tag{4}$$

$$\begin{aligned}\mathcal{L}_G = \mathcal{L}_{\text{GAN}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} \\ + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}}.\end{aligned} \tag{5}$$

CycleGAN Training Details

The CycleGAN model was trained using a ResNet-based generator with 9 residual blocks and a PatchGAN discriminator. All images were resized to 256×256 pixels. Training was performed for 20 epochs with a batch size of 1.

The Adam optimizer was used for both generators and discriminators with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate was set to 5×10^{-5} for all networks. The cycle-consistency loss weight was set to $\lambda_{cyc} = 10$, while the identity loss weight was set to $\lambda_{id} = 4$ with an additional scaling factor of 0.5. A VGG16-based perceptual loss was incorporated with a weighting factor of 0.4 to encourage high-level semantic consistency. The dataset was split into 70% training, 15% validation, and 15% testing sets.

Physiology-informed priors (skin flushing, gaze jitter, red-eye) are injected to improve realism and variability. This augmentation enriches the drunk class with physiologically meaningful features, thereby supporting more robust downstream classification. As illustrated in Figure. 2, the CycleGAN pipeline employs a ResNet-9 generator and a PatchGAN discriminator with multiple complementary losses.

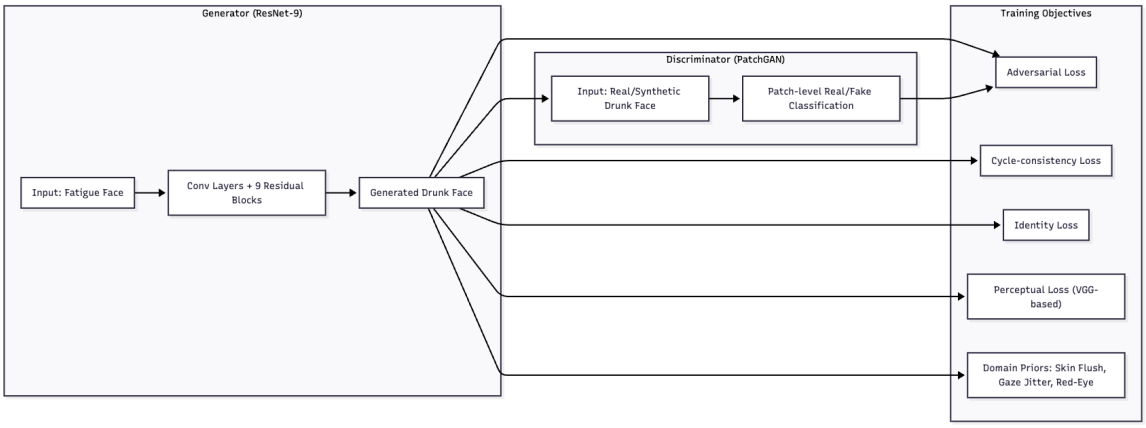


Figure 2: CycleGAN-based augmentation pipeline.

Representative outputs of the CycleGAN are shown in Figure. 3, illustrating how fatigue-domain inputs are translated into synthetic drunk faces that preserve identity while embedding alcohol-related cues such as skin flushing and gaze instability.



Figure 3: Example of a synthetic drunk face generated by CycleGAN from a fatigue-domain input.

3.4 SE-Enhanced MobileNetV2 Classifier

For multi-class recognition, this study adopts a MobileNetV2 backbone [15], chosen for its efficient inverted residual and linear bottleneck design. Early layers are frozen to preserve general low-level representations, while the final three convolutional stages are fine-tuned to capture task-specific impairment cues. This transfer learning strategy ensures efficiency while avoiding overfitting.

Squeeze-and-Excitation Block

To capture subtle impairment signals, a Squeeze-and-Excitation (SE) block [13] is placed after global average pooling. The operation consists of three stages:

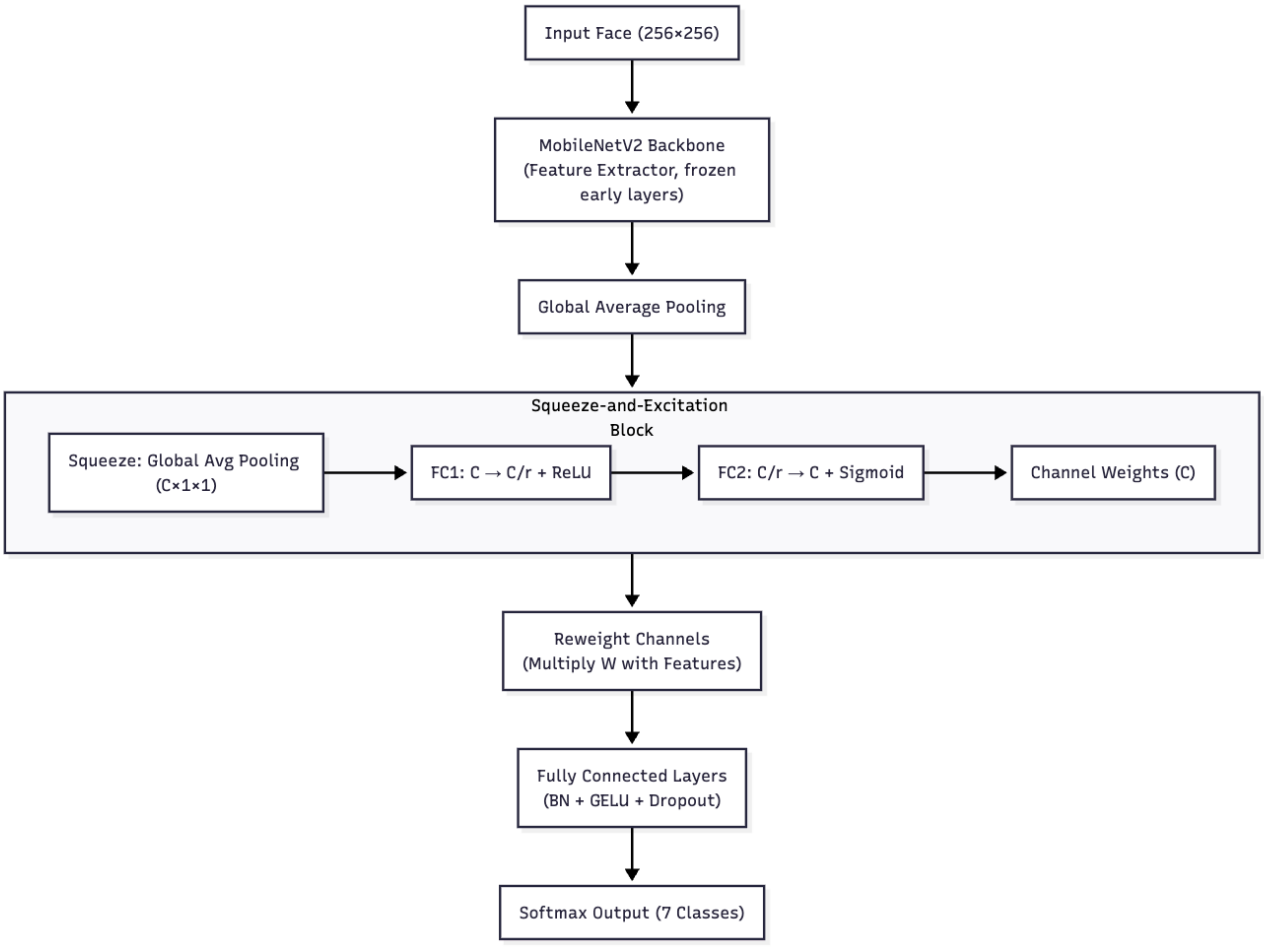


Figure 4: SE-Enhanced MobileNetV2 classifier. Early layers are frozen, later blocks fine-tuned. An SE block recalibrates channels post-GAP, and an MLP head with BN–GELU–Dropout maps features to seven driver states.

(i) Squeeze. Global average pooling aggregates spatial information across each channel into a descriptor vector $\mathbf{z} \in \mathbb{R}^C$, where C is the number of channels:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i, j).$$

(ii) Excitation. The descriptor is passed through two fully connected layers with a reduction ratio r, followed by ReLU and sigmoid activations to produce normalized channel weights:

$$\mathbf{s} = \sigma(W_2 \delta(W_1 \mathbf{z})), \tag{6}$$

where $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$, $\delta(\cdot)$ is ReLU, and $\sigma(\cdot)$ is sigmoid.

(iii) Recalibration. The weights $\mathbf{s}$ are applied channel-wise to the original feature maps:

$$\hat{\mathbf{X}} = \mathbf{s} \odot \mathbf{X}, \tag{7}$$

enhancing impairment-relevant channels (e.g., periocular redness, eyelid drooping, skin flushing) while suppressing irrelevant variations such as background or lighting.

Classifier Head

The classifier head consists of three fully connected layers with batch normalization, GELU activation, and dropout (rates 0.5, 0.3, 0.2). A final softmax layer maps features into seven driver states: safe, fatigue, drunk, drinking, hair and makeup, phone call, and talking to passenger.

This design provides three benefits: (i) MobileNetV2 ensures real-time feasibility. (ii) SE blocks model interchannel dependencies, amplifying subtle impairment cues. (iii) The multi-layer classifier head regularizes learning and improves generalization across diverse real-world conditions.

3.5 Optional Temporal Extension with LSTM

While the main focus of this work is frame-level recognition using a CycleGAN-augmented SE-MobileNetV2 backbone, a simple temporal extension was additionally explored to assess whether sequence modeling could further improve recognition. Specifically, frame-level features extracted from the MobileNetV2 backbone (1280d) were aggregated into fixed-length sequences of 30 frames and passed into a single-layer LSTM with hidden size 128 and dropout 0.3, followed by a fully connected classification layer. For each sequence, the ground-truth label was assigned by majority voting across its constituent frames. The LSTM baseline was trained using the Adam optimizer (lr=10^{-4}), cross-entropy loss, and a ReduceLROnPlateau scheduler.

4. RESULTS

4.1 GAN Training Stability

The CycleGAN was trained to generate synthetic drunk faces from fatigue-domain inputs. Figure 5 shows the generator loss ($\mathcal{L}_G$) across epochs for both training and validation. As seen in the plot, the loss converged smoothly with minimal divergence between training and validation curves, indicating stable adversarial training and reliable generation quality. This stability ensured that the augmented drunk samples could be integrated into the unified dataset without introducing artifacts or mode collapse.

4.2 Classifier Quantitative Performance

The SE-enhanced MobileNetV2 classifier was evaluated on the unified seven-class dataset. The model achieved an overall test accuracy of **97.67%** with a cross-entropy loss of **0.0655**. The detailed training and validation curves are presented in Figure. 6, confirming that the network converged stably without overfitting.

Table 1 summarizes the precision, recall, and F1-score for each category. The results show near-perfect performance for distraction-related behaviors (drinking, makeup, phone, passenger).

4.3 Temporal Baseline Results with LSTM

To complement the frame-based classifier, we report results of the MobileNetV2 feature extractor coupled with an LSTM sequence model. Although training curves indicated stable convergence, the sequence model plateaued at 63.01% test accuracy with cross-entropy loss of 1.0025, substantially lower than the 97.67% achieved by the frame-level SE-MobileNetV2. The class-wise report (Table 2) highlights strong performance in some categories (e.g., drunk F1=0.73, phone call F1=0.78) but near-zero recall for small classes such as hair and

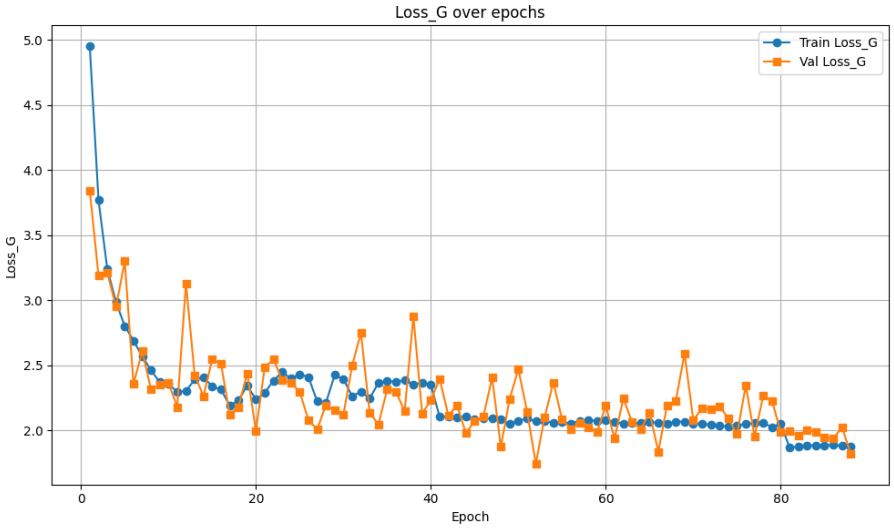


Figure 5: Generator loss curves ($\mathcal{L}_G$) for CycleGAN during training, demonstrating stable convergence across 90 epochs.

makeup or passenger talk. This suggests that, without sufficient temporal diversity, a simple recurrent baseline is not competitive for multi-class driver state recognition in our setting.

Figure 7 shows the accuracy and loss curves for the LSTM baseline. Training was stable, but validation curves plateaued early, further highlighting that the temporal recurrent extension without richer augmentation or advanced attention mechanisms could not match the frame-level SE-MobileNetV2 performance.

4.4 Real-World Qualitative Evaluation

Beyond numerical metrics, qualitative results provide further evidence of robustness. Figure 8 shows representative predictions across all seven driver states. Each sample demonstrates accurate recognition under varying conditions, including illumination changes, occlusion, and subtle impairment cues. The model consistently preserved discriminative ability across both distraction behaviors and synthetic drunk faces.

5. DISCUSSION

The proposed framework integrates generative augmentation with attention-enhanced lightweight CNN classification, achieving a test accuracy of **97.67%** across seven classes. This section compares the approach with recent driver monitoring methods and highlights its advantages as well as remaining challenges.

5.1 Comparative Analysis

Table 3 presents a comparison against representative models from the literature. Cañas et al. [18] investigated distraction recognition on the Vicomtech DMD dataset, achieving 97.3% using a MobileNet+LSTM video-based architecture. While effective for distraction, their method does not extend to fatigue or alcohol impairment. Baheti et al. [19] employed a VGG-based CNN on the StateFarm dataset, reporting 94–96% accuracy but again focusing solely on distraction. More recently, the EFFNet-CA architecture [20] introduced channel attention for distraction recognition, reaching 96.6%, yet remaining limited to single-task classification. Song et al. [21] developed CFNet, a CNN-fusion model for distraction, with 93.45% accuracy.

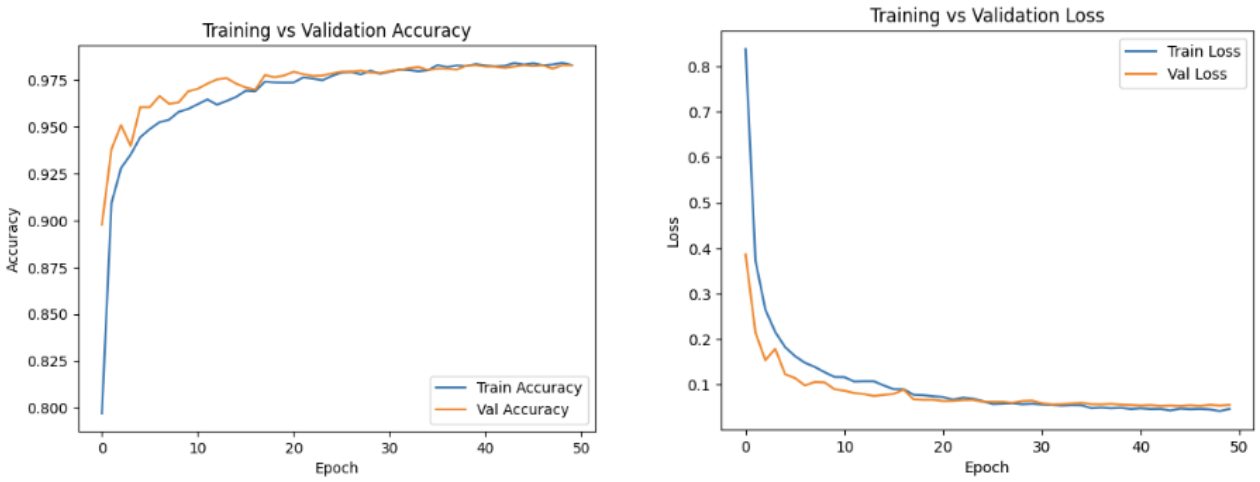


Figure 6: Performance curves (accuracy and loss) of the SE-enhanced MobileNetV2 classifier.

Table 1: Classification report on the test set.

Class	**Precision**	**Recall**	**F1-score**
Safe drive	0.91	0.94	0.92
Fatigue	0.97	0.93	0.95
Drunk	1.00	1.00	1.00
Drinking	1.00	1.00	1.00
Hair/Makeup	0.99	1.00	0.99
Phone Call	1.00	1.00	1.00
Passenger Talk	0.98	0.98	0.98
Overall Acc.		0.9767	
Macro Avg.	0.98	0.98	0.98
Weighted Avg.	0.98	0.98	0.98

Compared to these approaches, this study not only attains competitive or superior performance but also expands the recognition scope to include fatigue and alcohol impairment. The inclusion of synthetic drunk augmentation via CycleGAN directly addresses the long-standing issue of alcohol-related data scarcity, a dimension absent in the above works. Furthermore, the integration of Squeeze-and-Excitation (SE) attention improves the capture of subtle cues such as gaze instability and facial flushing, leading to more robust performance across diverse impairments.

5.2 Limitations

Despite the strong results, certain limitations remain. Alcohol impairment is not always manifested through facial cues, leading to possible false negatives. Moreover, synthetic drunk faces generated by CycleGAN, while diverse, cannot fully substitute for large-scale real-world data.

Table 2: Classification report of the LSTM temporal baseline.

Class	Precision	Recall	F1-score	Support
Safe drive	0.38	0.55	0.44	22
Fatigue	0.62	0.72	0.67	25
Drunk	0.78	0.69	0.73	26
Drinking	0.50	0.43	0.46	7
Hair / Makeup	0.00	0.00	0.00	7
Phone Call	0.73	0.84	0.78	49
Passenger Talk	0.00	0.00	0.00	10
Accuracy		0.63 (146 samples)		
Macro Avg.	0.43	0.46	0.44	146
Weighted Avg.	0.57	0.63	0.60	146

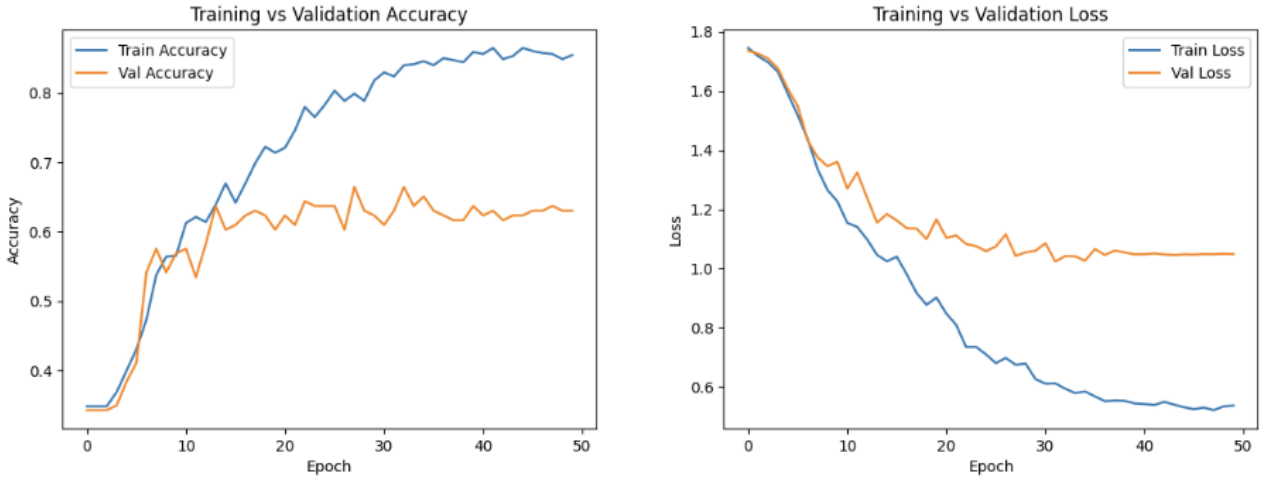


Figure 7: Performance curves (accuracy and loss) of LSTM temporal baseline.

6. CONCLUSION AND FUTURE WORK

This study presented a unified framework for driver state recognition that integrates CycleGAN-based synthetic augmentation with an SE-enhanced MobileNetV2 classifier. The approach achieved 97.67% accuracy and demonstrated resilience across fatigue, distraction, and drunk categories. By addressing dataset imbalance through generative modeling, the system improved recognition of conditions that are otherwise underrepresented in real data. Future work will focus on expanding real-world datasets, incorporating multimodal signals beyond facial appearance, and exploring temporal modeling to capture dynamic transitions between driver states. Deployment in diverse vehicle environments, including infrared or low-light conditions, will further enhance applicability for intelligent transportation systems.

REFERENCES

[1] L. Vacchetti, V. Lepetit, and P. Fua, "Stable real-time 3D tracking using online and offline information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 10, pp. 1385–1391, 2004.



Figure 8: Representative qualitative results of driver state recognition across seven categories. The model demonstrates robust performance under diverse conditions, confirming its ability to generalize to real-world scenarios.

[2] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[6] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa, "Real-time distracted driver posture classification," *arXiv preprint arXiv:1706.09498*, 2017.

[7] G. Bresson, M.-C. Rahal, D. Gruyer, M. Revilloud, and Z. Alsayed, "A cooperative fusion architecture for robust localization: Application to autonomous driving," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 2016, pp. 859–866.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[9] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

Table 3: Comparison with recent driver-monitoring methods.

Work	Model / Technique	Dataset / Task	Accuracy
Proposed (2025)	MobileNetV2 + SE + CycleGAN drunk augmentation	DMD + Kaggle Drowsiness + synthetic drunk (multi-class)	97.67%
Cañas et al. (2021) [18]	MobileNet + LSTM (video-based)	Vicomtech DMD (distraction only)	97.3%
Baheti et al. (2018) [19]	VGG-16 CNN with dropout	StateFarm (distraction only)	94.4–96.3%
Abbass et al. (2023) [20]	EFFNet-CA (MobileNet + channel attention)	SFD3, AUCD2 (distraction only)	96.6%
Song et al. (2023) [21]	CF-Net (CNN feature fusion)	SynDD1 (distraction)	93.45%

[10] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[11] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[12] D. He, L. Yang, X. Tian, and Z. Wang, "An overlapped decomposition optimization method for dynamic economic dispatch," *IEEE Access*, vol. 6, pp. 45804–45820, 2018.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[16] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[18] P. Cañas, J. Ortega, M. Nieto, and O. Otaegui, "Detection of distraction-related actions on dmd: An image and a video-based approach comparison," 01 2021, pp. 458–465.

[19] N. Darapaneni, J. Arora, M. Hazra, N. Vig, S. S. Gandhi, S. Gupta, and A. R. Paduri, "Detection of distracted driver using convolution neural network," *arXiv preprint arXiv:2204.03371*, 2022.

[20] T. Khan, S. Anwar et al., "Effnet-ca: An efficient driver distraction detection based on multiscale features extractions and channel attention mechanism," *Sensors*, vol. 23, no. 8, p. 3835, 2023.

[21] L. Song, H. Cui, X. Zhao, and M. Sun, "Distracted driving behavior detection based on cnn and multi-model fusion," in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 12636, 2023, p. 126363Z.