# ETANet: EfficientNet-Transformer Attention Network for Robust Driver Drowsiness Detection

***Abstract*—Driver drowsiness is a major contributing factor to road accidents worldwide, motivating the need for accurate and real-time monitoring systems. While convolutional neural network (CNN)-based approaches have achieved strong performance in controlled settings, they often struggle to generalise across variations in lighting, occlusion, and subject appearance.**

**In this work, we evaluate a CNN-based baseline model and propose an EfficientNet–Transformer attention network, ETANet that combines EfficientNetB0 for spatial feature extraction with a Transformer encoder to model global spatial feature interactions. The goal is to investigate whether incorporating self-attention mechanisms improves robustness for drowsiness detection.**

**Experiments are conducted on the MRL Eyes dataset and the NTHU-DDD dataset to evaluate both in-domain performance and cross-dataset generalization. Models are compared using accuracy, precision, recall, F1-score, and AUC.**

**The proposed model achieves strong performance, reaching 99.05% accuracy on the MRL Eyes dataset and 99.04% accuracy on the NTHU-DDD dataset, substantially outperforming CNN-based baselines. These results suggest that combining convolutional feature extraction with attention-based modeling can improve robustness in vision-based driver monitoring systems.**

***Index Terms*—ETANet, Transformers, Deep Learning, CNN, Driver Drowsiness Detection**

## I. Introduction

Driver drowsiness is one of the leading causes of road accidents globally. Such risks necessitate the development of a real-time detection system in vehicles that checks for driver drowsiness and alerts the driver with a warning if it detects they are not alert. Fatigue significantly reduces a driver's reaction time, cognitive processing speed, and situational awareness. While deep learning to detect facial and eye regions for driver alertness has advanced, convolutional neural networks (CNNs) often struggle with capturing long-range spatial dependencies. Drowsiness is inherently a temporal problem characterized by subtle behaviors that include prolonged eye closure, slow blinking, and micro-sleeps. Hence, robustness at the frame-level is needed when temporal modeling is explicitly unavailable.

Transformers mitigate the locality bias of standard CNNs as they utilize self-attention to weigh the importance of different features globally. This research proposes an EfficientNet-Transformer Attention Network architecture (ETANet). The model aims to capture global dependencies across facial regions with its attention mechanism. The aim of this research is to improve classification performance in transitional fatigue states. The secondary aim of this research is to develop a highly accurate yet robust system that is generalizable across diverse data and environmental conditions.

The contributions of this paper are as follows:

- **ETANet Architecture Design:** We investigate ETANet, which combines EfficientNetB0 feature extraction with a Transformer encoder to incorporate global feature interactions for drowsiness detection.
- **Cross-Dataset Evaluation:** We evaluate model performance on both the MRL Eyes dataset and the NTHU-DDD dataset to assess robustness under different environmental and data distribution conditions.
- **Real-Time Feasibility Analysis:** We evaluate inference latency under CPU-based execution to assess the suitability of the proposed approach for real-time driver monitoring applications.

## II. Background

Recent studies on driver drowsiness detection have been extensively studied in a systematic review by Fonseca and Ferreira [1], who analyzed over 80 deep-learning-based approaches from 2015 to 2025. Their findings show that most CNNs extract facial features such as eye closure and yawning. However, they also found a shift towards multi-modal datasets that combine facial, physiological, and vehicle-based signals. They also found a gradual transition towards transformer and hybrid architectures. The review identifies that limitations in demographic bias and poor generalization to unseen environments persist [1].

With regard to CNN-based architectures. MobileNet is a widely adopted, lightweight, and effective option that has achieved high accuracy and outperformed heavier CNNs like VGG16 [6]. Similarly, Bhanja et al. eliminated false alarms using MobileNet, although at the expense of recall [8]. While these studies show the practicality of CNNs, they also highlight their limitations to occlusions and light variation.

A hybrid CNN-based model approach, EffRes-DrowsyNet [4], combines EfficientNetB0 and ResNet50 to push accuracy on the NTHU-DDD dataset [9] to 97.7%. However, the model is limited by its reliance on frame-level analysis, "black-box" nature, and high computational intensity. Multi-modal neural networks like those proposed by Cao et al. [2] integrate facial, physiological, and vehicle signals using attention mechanisms and enable dynamic weighting of modalities depending on the environment. Hassan et al. [3] achieved high accuracy with real-world deployment constraints taken into account. However, these systems' dependence on complex multi-modal data limits their scalability to be deployed in real-world driving environments.

More recent work has begun exploring transformer-based architectures for drowsiness detection [3]. This method uses self-

attention mechanisms to capture global dependencies across facial regions. By leveraging self-attention, the model is able to capture subtle global spatial relationships that are associated with fatigue. Examples of this are prolonged eye closure, slow blinking and micro-sleep. Unlike multi-modal systems that rely on physiological sensors, vision-only Transformer models remain deployable using standard in-cabin cameras.

Moreover, unlike traditional CNN models like MobileNetV2, which focus on local receptive fields, Transformer-based models capture global dependencies [3] [7]. Consequently, recent literature emphasizes the need for models that are accurate in varying lighting conditions, occlusions like sunglasses, and demographically diverse data. This work builds upon these foundations by utilizing attention-based modeling with Transformers, which provides superior performance.

## III. Methodology

This study has a progressive pipeline consisting of (i) reproduction of a MobileNet model, (ii) evaluation under identical training and pre-processing conditions, and (iii) comparison against the proposed ETANet model. This design ensures that observed performance gains can be attributed to architectural differences rather than data handling or training protocol.

### A. Data Preprocessing

Images were resized to a consistent resolution for each experiment, and grayscale datasets were represented as 3-channel inputs where required by pretrained backbones such as EfficientNetB0 and MobileNetV2. Pixel values are normalized to the range [0, 1]. Data augmentation was applied uniformly and includes random horizontal flips, random rotations (±5 degrees), and random zoom (±10%) to improve robustness to environmental variation.

### B. Baseline Model Reproduction

The baseline model is reproduced following the original MobileNetV2-based architecture and training configuration as closely as possible [5]. All preprocessing steps are kept consistent across models wherever applicable to ensure fair comparison. Any differences in input resolution or augmentation settings are explicitly reported in later sections.

The baseline model uses:

- **Input image resolution:** 224 x 224
- **Architecture:** MobileNet
- **Optimizer:** Adam Optimizer
- **Learning rate:** 0.001
- **Loss function:** Mean squared error
- **Batch size:** 32

The data was split into training, validation, and test sets following the original protocol. Although binary cross-entropy is more commonly used for binary classification, mean squared error was retained in the baseline reproduction to ensure strict adherence to the original implementation and allow for fair comparison between architectures. The only implementation-level modification was the use of a data streaming pipeline instead of loading the full dataset into memory at once. This change does not affect the optimization objective, gradient composition, or data distribution. It is mathematically equivalent to the array-based training. The modification was introduced for memory efficiency and does not affect model behavior or learning dynamics.

### C. Proposed CNN-Transformer Architecture

This paper presents an EfficientNet–Transformer attention network (ETANet), which augments the EfficientNetB0 model with a feature-level self-attention encoder in order to capture global spatial dependencies within the facial features. The proposed model builds upon the limitations observed in CNN-based baseline reproduction and failure of the fine-tuned model, and integrates EfficientNetB0 with a Transformer encoder to combine the spatial feature extraction of a CNN with the global relationship modeling of an attention-based Transformer. The ETANet architecture is shown in Figure 1.

- **Input:** The input consists of grayscale facial images resized to 224 x 224 pixels. The input images are grayscale drowsiness images. Although the source images are grayscale, they are represented as 3-channel inputs to match the ImageNet-pretrained EfficientNetB0 backbone.
- **Feature Extraction:** An EfficientNetB0 backbone, pre-trained on ImageNet, is used to extract high-level spatial features from input images. The model replaces earlier CNN backbones, such as MobileNetV2 used in the baseline and fine-tuned experiments, due to its strong performance-efficiency trade-off in vision tasks.
- **Feature Tokenization:** The extracted feature maps are flattened and converted into sequential tokens. Positional encoding is utilized to maintain spatial relationships between facial regions, allowing the model to understand structural dependencies.
- **Transformer Encoder:** Multiple Transformer encoder layers apply Multi-Head Self-Attention (MHSA) and Feed-Forward Networks (FFN). These are applied to capture global dependencies across the features previously tokenized. The model is able to capture dependencies like simultaneous eye closure and yawning patterns associated with drowsiness.
- **Classification Head:** The final output is aggregated using Global Average Pooling and a Multi-Layer Perceptron (MLP). The output is a probability score produced by the sigmoid activation function to perform binary classification to predict whether a driver is drowsy or alert.

Generalization is improved by the use of label smoothing, which prevents overconfidence. Dropout (0.2) with weight decay is used for regularization. The model is trained using a two-stage strategy to ensure stable learning and improved convergence behaviour. The model is evaluated against a baseline MobileNetV2 and a fine-tuned MobileNetV2 variant. Performance is assessed using accuracy, Area Under the Curve (AUC), F1-score, precision, and recall.

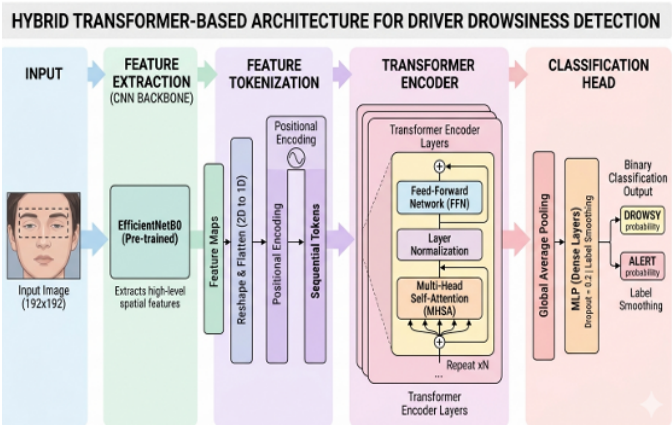


Fig. 1. ETANet Architecture for Driver Drowsiness Detection

## IV. Experimental Setup

The experimental pipeline consists of three stages: baseline reproduction, CNN fine-tuning, and ETANet based evaluation across two datasets. To ensure fair comparison, all models are evaluated under consistent preprocessing and training protocols unless explicitly stated otherwise.

**Baseline Reproduction:** The original baseline model uses MobileNetV2 as a lightweight CNN architecture pretrained on ImageNet. This choice was made for computational efficiency and suitability for real-time drowsiness detection. The reproduced MobileNetV2 model serves as the primary baseline for comparison against other approaches, including the fine-tuned MobileNetV2 variant and the proposed ETANet model.

**Fine-Tuned CNN:** The MobileNetV2 model is further fine-tuned on the target dataset to improve task-specific performance and reduce domain mismatch. For this, the pretrained baseline model was loaded and adapted to the new drowsiness dataset. During fine-tuning, the model became biased toward predicting the drowsy class, as shown by its recall of 1.0000 but low accuracy and precision. This indicates class collapse, where the model learned a trivial prediction strategy instead of separating drowsy and non-drowsy samples reliably.

**ETANet Model:** The proposed ETANet model is trained and evaluated both on the original dataset (MRL) and a separate unseen dataset (NTHU-DDD) to test generalization performance.

### A. Datasets and Evaluation Metrics

Experiments were conducted on two datasets

- **MRL Eyes 2018 Dataset:** Used for baseline reproduction and initial controlled experiments. It contains over 50,000 labeled images of eye states under relatively clean conditions.
- **NTHU-DDD Dataset:** A large-scale benchmark dataset with over 66,000 images containing variations in lighting, head pose, occlusions (e.g., glasses), and real driving conditions.

All models are trained using the Adam optimizer with binary cross-entropy loss. Data augmentation includes random horizontal flips, rotations (±5%), and zoom (±10%) to improve robustness.

## V. Results

The models were evaluated using accuracy, AUC (Area Under Curve), F1-score, confusion matrix precision, and recall. The complete results of the evaluation metrics are presented in the Table I. The ETANet model achieved an accuracy of 0.9905 on the Media Research Lab (MRL) Eyes dataset, and achieved an accuracy of 0.9904 on the NTHU-DDD dataset, while balancing precision and recall, and effectively avoiding the class collapse observed in the fine-tuned CNN. The fine-tuned CNN had the lowest performance with an accuracy and precision of 0.549. Its recall was 1, showing class collapse. Compared to the baseline CNN, ETANet showed increased accuracy, recall, and F1 score.

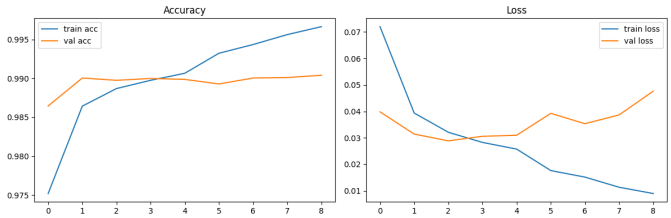


Fig. 2. ETANet accuracy on (old) MRL Eyes Dataset

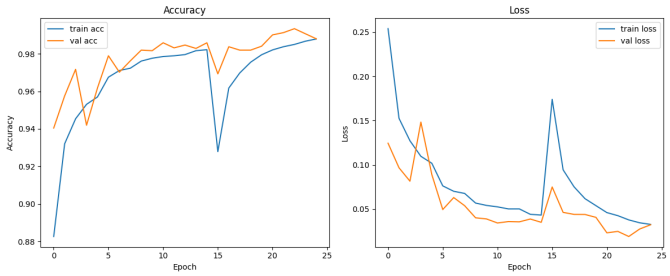


Fig. 3. ETANet accuracy NTHU-DDD (new) Datasets

| | Baseline Reproduction | ETANet MRL Eyes | ETANet NTHU-DDD | Fine-tuned |
|---|---|---|---|---|
| Accuracy | 0.8493 | 0.9905 | 0.9904 | 0.5499 |
| Precision | 0.8640 | 0.9929 | 0.9946 | 0.5499 |
| Recall | 0.8334 | 0.9878 | 0.9840 | 1.0000 |
| F1 | 0.8484 | 0.9904 | 0.9892 | 0.7096 |
| AUC | 0.9241 | 0.9998 | 0.9997 | - |

TABLE I
Evaluation Metrics to compare Baseline, ETANet, and Fine-tuned CNN performance

The Figures 2 and 3 show the high performance of the ETANet model on both the MRL Eyes Dataset and the NTHU-DDD Dataset. The figure shows a comparison between training and validation accuracy and loss for the model with both datasets.

For the new drowsiness dataset, ETANet achieved above 99 percent test accuracy. In the MRL experiment, early stopping

restored the best Stage 2 weights from epoch 4, while the new-data model completed two-stage training with checkpoints and early stopping to select the best validation-AUC model.

The system was tested in real-time using a webcam. The system achieved an inference latency of 80-120 ms per frame on CPU hardware, showcasing its robustness and feasibility for use in real-time systems with minimal delays.

## VI. Ablation Study

To quantify the specific contribution of the Transformer encoder, we conducted an ablation study comparing the full architecture against a baseline consisting solely of the EfficientNetB0 backbone.

### A. Setup

In the main proposed model, the pipeline is: EfficientNetB0 is used to extract feature maps, which are tokenized and fed to the Transformer encoder. Finally, the classification head produces output; however, in an ablation study, the Transformer part is removed. Both models still use EfficientNetB0 and are trained under the same dataset setup, but only the proposed model has self-attention. This makes the comparison fair because the main architectural difference is the Transformer block.

### B. Results

The ablation result shows that EfficientNetB0 alone performs well, however it is consistently outperformed by the full ETANet model as shown in Table II. On the NTHU-DDD dataset, the EfficientNet-only variant achieved about 95.37% accuracy, while ETANet achieved 99.04% accuracy. This improvement of around 3.67 percentage points suggests that the Transformer encoder helps the model capture relationships between spatial features, such as patterns around the eyes and face, better than the CNN backbone alone. There is also a notable reduction in false negatives from 152 to 24, this shows that ETANet is better at identifying positive cases where the subject was drowsy. The ablation study demonstrates that the Transformer encoder within ETANet improves performance and reduces missed detections.

| | **ETANet Model (EffNetB0+Trans)** | **Ablated Model (EffNetB0 Only)** |
|---|---|---|
| Accuracy | 0.9904 | 0.9537 |
| Precision | 0.9946 | 0.9985 |
| Recall | 0.9839 | 0.8985 |
| F1 Score | 0.9892 | 0.9459 |
| Confusion Matrix | TN=1815, FP=8<br>FN=24, TP=1481 | TN=1829, FP=2<br>FN=152, TP=1345 |

TABLE II

Ablation results for the proposed architecture comparing the full Transformer model against the EfficientNetB0 baseline

## VII. Discussion

The experimental results demonstrate that model architecture plays a critical role in the ability to generalize drowsiness detection across different data distributions. The experimental results suggest that augmenting convolutional feature extractors with a Transformer encoder can improve performance and generalization under dataset shift.

This improvement may be attributed to the model's ability to capture higher-order interactions between spatial regions of the face that are not explicitly modeled by convolutional filters alone.

In contrast, the poor performance of the fine-tuned MobileNetV2 model indicates that standard transfer learning approaches are highly sensitive to domain shifts, leading to a degenerative decision boundary under a fixed classification threshold. Although the model performs well on data similar to its training distribution, its inability to adapt to the new dataset suggests that CNN-based architectures may overfit to dataset-specific visual patterns rather than learning robust fatigue-related representations.

Another key observation is the trade-off between performance and computational complexity. While the Transformer model substantially improves classification performance and generalization, it requires more computational resources due to the self-attention mechanism and increased model depth. Despite the fact that the measured inference latency of approximately 80–120 ms per frame demonstrates that the model remains suitable for real-time applications, particularly in driver monitoring systems, where slight delays are acceptable.

From a safety perspective, the reduction in false negatives is particularly important, as failing to detect a drowsy driver poses a higher risk than occasional false alarms. The ETANet's balanced precision and recall suggest improved reliability in practical deployment scenarios compared to CNN-based baselines, which tend to exhibit biased or unstable predictions under distribution shifts.

However, the study is still limited by its focus on image-based inputs alone. Real-world driver drowsiness is a temporal phenomenon, and future improvements could involve explicitly modeling sequential frame information to further enhance detection stability over time. Additionally, further testing across more diverse demographic and environmental conditions would strengthen the robustness claims of the proposed approach.

## VIII. Conclusion

This research successfully demonstrates the efficacy of ETANet for robust driver drowsiness detection. Key findings indicate that while CNN baselines perform well on static datasets, they struggle to generalize to new environments, whereas ETANet maintains high accuracy and robust performance achieving 99.04% on the NTHU-DDD dataset and 99.05% on the MRL Eyes dataset. The integration of attention mechanisms provides a pathway for more reliable Driver Monitoring Systems, ultimately contributing to enhanced road safety and fewer accidents. Future work should focus on further optimizing these architectures for deployment in highly resource-constrained embedded environments.

## REFERENCES


[1] T. Fonseca and S. Ferreira, "Drowsiness Detection in Drivers: A Systematic Review of Deep Learning-Based Models," *Appl. Sci.*, vol. 15, no. 16, p. 9018, Aug. 2025. doi: 10.3390/app15169018.

[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, Apr. 2025. doi: 10.1038/s41598-025-86709-1.

[3] 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, May 2025. doi: 10.1038/s41598-025-02111-x.

[4] 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, pp. 3711–3711, Jun. 2025. doi: 10.3390/s25123711.

[5] H. A. Hussain, M. A. Subhi, A. S. Al, A. D. Radhi, and M. A. Albahar, "Improving Road Safety Through Multimodal Deep Learning for Driver Drowsiness Detection," *Iraqi J. Comput. Sci. Math.*, 2025. doi: 10.52866/2788-7421.1384.

[6] T. S. Delwar *et al.*, "AI- and Deep Learning-Powered Driver Drowsiness Detection Method Using Facial Analysis," *Appl. Sci.*, vol. 15, no. 3, pp. 1102–1102, Jan. 2025. doi: 10.3390/app15031102.

[7] S. V. Dixith, S. Jadhav, Y. Kim, and N. Jayakumar, "Real-Time Driver Drowsiness Detection," in *Proc. IEEE*, 2024. [Online]. Available: https://ieee.org. Accessed: Feb. 27, 2026.

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

[9] "NTHU-DDD: Driver Drowsiness Dataset," Emergent Mind, 2025. [Online]. Available: https://www.emergentmind.com/topics/nthu-ddd-dataset. Accessed: Feb. 27, 2026.