IJWR

Multi-Modal Driver Drowsiness Detection in ADAS via Attention-Guided Siamese Network with Temporal Modeling

Mahdi Seyfipoor*[a], Mohadese Parvizi[b], Siamak Mohammadi[a]

[a] Department of Computer Engineering, School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran; mahdiseyfipoor@ut.ac.ir, smohamadi@ut.ac.ir

[b] Department of Computer Engineering, Faculty of Engineering, Alzahra University, Tehran, Iran; ParviziiMohadese@gmail.com

ABSTRACT

Driver drowsiness detection plays a critical role in improving road safety, as drowsiness substantially increases the likelihood of traffic accidents. In this study, we propose a novel multi-modal framework within Advanced Driver Assistance Systems (ADAS) that leverages an Attention-Guided Siamese Network coupled with temporal modeling to accurately capture both spatial and temporal patterns of driver fatigue. The Siamese network processes paired facial images, enabling the extraction of discriminative features that highlight subtle changes in driver state. The attention mechanism is explicitly applied to the spatial feature maps within each branch of the Siamese network, allowing the model to focus selectively on key facial regions—such as eyes and mouth—that are most indicative of drowsiness, while also weighting complementary sensor modalities dynamically. Temporal modeling is incorporated through a sequential module (e.g., LSTM or temporal convolution) that analyzes the extracted features over time, capturing gradual and evolving signs of drowsiness that static frame-based methods often overlook. Extensive evaluations on benchmark datasets (YawDD, NTHUDDD) and a novel real-world driving dataset demonstrate superior accuracy exceeding 98.8%, along with strong cross-subject generalization. Ablation studies confirm the critical contributions of the attention mechanism in improving feature discrimination, and the temporal modeling module in enhancing sensitivity to progressive drowsiness. The proposed method surpasses traditional approaches in temporal awareness, data efficiency, and resilience to inter-subject and environmental variations, offering a robust and interpretable solution for real-time driver drowsiness monitoring in intelligent vehicles.

Keywords— Driver Drowsiness Detection, Multi-modal Fusion, Attention-Guided Siamese Network, Temporal Modeling, ADAS, Contrastive Learning, Facial Feature.

1. Introduction

Driver Drowsiness remains a significant challenge in road safety, contributing to a substantial number of traffic accidents each year. Advanced Driver Assistance Systems (ADAS) represent a backbone of modern vehicle safety, designed to enhance driving experience while mitigating risks on the road. These systems function by assisting drivers in various aspects of vehicle control, such as maintaining safe distances, monitoring blind spots, and detecting driver states like distraction and drowsiness. Despite significant advancements, road accidents continue to present a major threat to public safety, demonstrating the need for improvement in existing detection and intervention protocols. Among the many causes of road accidents, driver drowsiness, distractions, and abnormal behaviors account for a substantial proportion of incidents, often leading to catastrophic outcomes. According to the National Highway Traffic Safety Administration (NHTSA) [1], drowsy driving is directly responsible for tens of thousands of traffic accidents annually, with fatigue and inattention impairing reaction times, decision making, and general situational awareness. In response to this challenge, ADAS technologies increasingly emphasize the integration of driver behavior monitoring systems into vehicle safety frameworks to analyze and classify driver

http://dx.doi.org/10.22133/ijwr.2025.525802.1289

Citation M. Seyfipoor, M. Parvizi, S. Mohammadi, " Multi-Modal Driver Drowsiness Detection in ADAS via Attention-Guided Siamese Network with Temporal Modeling", *International Journal of Web Research*, vol.9, no.1,pp.39-56, 2026, doi: http://dx.doi.org/10.22133/ijwr.2025.525802.1289.

*Coressponding Author

Article History: Received: 24 May 2025; Revised: 22 November 2025; Accepted: 29 November 2025.

alertness levels in real time. These systems aim to alert drivers using various mechanisms—such as vibration feedback or auditory warning sounds whenever abnormal behavior, such as drowsiness or inattention, is detected. A critical element of ADAS is its ability to evaluate driver states, including drowsiness, exhaustion, and other unsafe behaviors [2]. Identifying signs of fatigue involves monitoring both physiological and behavioral indicators that signal a driver's diminishing control and focus. Indicators such as yawning, prolonged eye closure, reduced blink rates, erratic head movements, and difficulty maintaining consistent lane position highlight the onset of fatigue-related driving behaviors. Once detected, ADAS systems aim to mitigate accident risks by employing alerts, automation, or even direct intervention mechanisms, such as braking assistance or steering correction [3]. Traditional detection methods have primarily relied on physical signals such as heart rate variability, electroencephalograms (EEG) [4], and electrocardiograms (ECG) [5]. In recent years, researchers and engineers have turned to non-intrusive computer vision-based techniques as a solution to the shortcomings of traditional methods. By leveraging advancements in artificial intelligence (AI), machine learning (ML), and deep learning algorithms, modern ADAS systems can detect and analyze driver behaviors through real-time visual monitoring. These systems focus on facial expressions, eye blink rates, yawning frequency, and head tilt to identify potential signs of fatigue. Unlike physiological techniques, such approaches rely on onboard cameras and algorithms powered by convolutional neural networks (CNNs). This integration of AI and vision enables real-time monitoring without requiring the driver to wear or use external sensors, making the technology more practical and scalable across diverse vehicle types and demographics [6]. Unlike traditional CNN-based models, which focus on classifying individual images, (Siamese Neural Networks) SNNs excel in detecting subtle temporal and spatial variations by comparing pairs of input data. Specifically, this framework captures a baseline of the driver's alert facial expressions—such as eye aspect ratio (EAR), mouth aspect ratio (MAR), and head position—during initial calibration. Any subsequent deviations from this baseline are analyzed to detect drowsiness dynamically. The architecture comprises twin subnetworks sharing weights for feature extraction. Each subnetwork processes either baseline or real-time facial data to draw comparisons, allowing the detection of nuanced micro-changes in driver behaviors. Relative change learning, an innovative element of this approach, uses cosine similarity and contrastive loss to fine-tune detection thresholds dynamically. Figure 1 shows a view of the driver drowsiness detection system in an ADAS.

In recent years, the demand for innovative approaches in driver drowsiness detection has grown substantially, largely due to the critical need to improve road safety and mitigate accidents caused by driver fatigue. Conventional unimodal detection systems, which depend on a single source of information such as facial cues, eye movements, or vehicle dynamics, often struggle to maintain reliable performance in real-world scenarios characterized by diverse driving behaviors, lighting conditions, and environmental disturbances. Consequently, multi-modal frameworks have gained increasing attention, as they combine complementary data sources to achieve more robust, adaptive, and context-aware drowsiness detection.

Moreover, recent progress in deep learning has played a pivotal role in advancing these systems. In particular, the incorporation of attention mechanisms allows models to selectively focus on the most informative features across different modalities, while Siamese network architectures enable effective learning of similarity and temporal changes in driver behavior.

These methods not only allow for effective temporal modeling of driver behavior but also support personalized detection by comparing an individual's current state to their baseline patterns. Moreover, the shift toward real-time ADAS applications and the growing complexity of human-machine interaction in semi-autonomous vehicles necessitate innovative solutions capable of dynamic adaptation and rapid decision-making. Collectively, these factors underscore the need for novel, intelligent frameworks in the field of driver state monitoring.

Recent research by Cao et al. has introduced a multimodal neural network approach combining facial and physiological signals under sleep-deprivation conditions, demonstrating improved accuracy in real-world fatigue detection scenarios [7]. Qian et al. proposed a novel method based on

Figure. 1. Driver drowsiness detection in ADAS

facial landmark detection and temporal sequence transformers to better capture the dynamic characteristics of driver drowsiness [8]. To address the challenge of varying lighting conditions, a multimodal system using both visible and infrared imagery with a Bi-LSTM architecture and multi-loss fusion has been developed, showing strong performance across day and night settings [9]. A recent approach called the Multi-Aware Graph Convolutional Network (MAGCN) tackles the limitations of spatial feature locality and inter-driver variability by introducing attention-aware and temporal-aware extractors [10].

Furthermore, privacy-aware models such as [11] using federated learning and spatial self-attention have shown promise in achieving robust drowsiness detection while preserving user data confidentiality.

Advantages of proposed system compared to existing machine learning techniques [11-13] for drowsiness detection is shown in Table 1. SiamEEGNet achieved superior cross subject performance (EEG data) by focusing on relative changes, a strategy adaptable to facial analysis [12].

YOLO based facial detectors achieved 100% precision in drowsiness contexts [14], suggesting Siamese networks could integrate similar lightweight models. Self-supervised Siamese variants like PSNSleep achieved 80% accuracy in sleep staging without negative samples [11], a method transferable to facial drowsiness detection.

Table 1. Advantages of Proposed System Compared to Existing Machine Learning and Deep Learning Techniques for Drowsiness Detection

Feature	***Traditional ML (SVM)***	***Deep Learning (YOLO-CNN)***	***Siamese Network***
Data Efficiency	Requires large labeled datasets	Needs extensive annotations	Learns from relative changes, reducing labeled data
Temporal Sensitivity	Static frame analysis	Limited temporal context	Captures drowsiness via baseline comparison
Cross-Subject Robustness	Manual feature assignment	Subject-specific tuning	Generalizes better via relative metrics
Interpretability	Black-box decisions	Complex feature maps	Aligns with clinically facial descriptors (EAR, MAR)

The future of ADAS lies in its ability to incorporate multimodal, intelligent, and adaptive frameworks for real-time safety monitoring. Early stopping is a regularization technique that halts model training when validation performance stabilizes or degrades, preventing overfitting while maintaining generalization capability. This method strategically balances model complexity by leveraging validation metrics as stopping criteria rather than arbitrary epoch counts. For drowsiness detection tasks, where models need to generalize across diverse drivers and environmental conditions, early stopping ensures that the neural network captures underlying patterns without overfitting to specific scenarios. This balance is critical for real-time applications where robustness and reliability are paramount. By employing early stopping, we optimized the neural network's ability to generalize effectively while conserving computational resources and avoiding unnecessary complexity in model architecture. This article is an extended version of [14].

2. Related Work

Recent studies on driver drowsiness detection have increasingly focused on multi-modal approaches combining facial features and other physiological signals to improve robustness and accuracy. Zhang et al. proposed a deep convolutional neural network (CNN) based model that leverages facial subsampling techniques to effectively detect signs of fatigue, achieving high accuracy under varying conditions [15]. Similarly, Wang et al. introduced a multi-source fusion method that integrates physiological signals such as heart rate with facial images, demonstrating enhanced performance over unimodal systems [16]. Liu et al. conducted a feasibility study on multi-modal driver drowsiness detection, combining facial expressions, body movements, and physiological data, which highlighted the potential of multi-modal systems to better capture the complexity of driver states [17]. Additionally, Jones reviewed AI-enabled smart cameras capable of monitoring driver behavior including eye blinking, head movement, and posture, showing promising results in real-time drowsiness alert systems [18]. These studies underscore the importance of multi-modal data fusion and advanced deep learning architectures to address the challenges of driver variability and environmental factors in drowsiness detection.

Recent advancements in driver drowsiness detection have increasingly integrated attention mechanisms and temporal modeling to enhance accuracy and robustness. For instance, Qian et al. proposed a fatigue detection method utilizing facial landmark localization and temporal sequence Transformers, addressing challenges like low real-

time performance and high false positive rates in existing deep learning-based methods [19].

In another study, Jin and Dong introduced YOLO11-CR, a lightweight convolution-and-attention framework for accurate fatigue driving detection. This model combines convolutional neural networks with attention mechanisms to improve feature extraction and classification performance [20]. Ren et al. developed LiteFat, a lightweight spatio-temporal graph learning model for real-time driver fatigue detection. By converting streaming video data into spatio-temporal graphs and employing a graph neural network, LiteFat efficiently captures temporal dependencies while maintaining low computational complexity [21].

Furthermore, Qu et al. proposed a multi-task learning approach for fatigue detection and face recognition of drivers via a tree-style space-channel attention fusion network. This model simultaneously performs fatigue detection and face recognition, enhancing both accuracy and efficiency [22]. These studies underscore the significance of integrating attention mechanisms and temporal modeling in multi-modal systems for effective driver drowsiness detection.

Detecting driver drowsiness has posed a long-standing challenge for researchers, particularly in the development of ADAS. With increasing expectations for safer transportation and the growing emergence of autonomous and semi-autonomous vehicles, the demand for real-time, accurate, and efficient driver monitoring systems increasing. Researchers have developed a diverse range of methodologies for drowsiness detection, spanning from physiological monitoring to computer vision-based systems and, more recently, hybrid artificial intelligence (AI) and deep learning models. These advancements have revolutionized the field, enabling faster, more efficient, and cost-effective solutions that identify early signs of driver fatigue, distraction, and dangerous driving behaviors.

The dominant approaches for assessing driver drowsiness fall into four primary categories, each offering distinct advantages and facing unique limitations.

2.1. Behaviral Monitoring via Computer Vision

Behavioral monitoring focuses on detecting visual signs associated with fatigue, often centered on the analysis of facial expressions and head movements. Among the most widely used techniques in this category are as follows.

Eye State Analysis

Methods that track and analyze the eye aspect ratio (EAR) and blink frequency to detect eyelid closure patterns. Modern approaches such as YOLO-based neural networks track eyelids with remarkable precision, achieving 100% accuracy in controlled settings. These systems use adaptive EAR thresholds to avoid misclassifying natural blinks as indicators of drowsiness [23] [24].

Mouth Movements

Yawning detection has been effectively implemented using mouth aspect ratio (MAR) measurements, combined with hybrid machine learning models like CNN-Haar cascades, achieving over 97% accuracy in distinguishing yawning gestures from regular speech [23].

Head Pose Estimation

Driver behavior is also characterized by subtle head nodding or tilting. Algorithms that analyze Euler angle deviations have reached 96.54% accuracy in simulated driving conditions, further enhancing their applicability [24]. While these techniques provide non-intrusive and accurate solutions, their performance often declines in complex, uncontrolled environments, such as during nighttime driving or when the driver is wearing glasses or masks.

2.2. Physiological Sensing

Physiological sensing relies on measurements of internal bodily signals to detect signs of drowsiness. These signals are highly correlated with fatigue levels and provide reliable markers for early-stage drowsiness detection. Key methodologies include:

EEG-Based Systems

Spectral power in specific brainwave frequency bands, such as delta (0.5–4 Hz), theta (4–8 Hz), and alpha (8–13 Hz), is analyzed using EEG sensors. Studies have shown that EEG-based systems can achieve 94% accuracy in fatigue classification, with prefrontal cortex channels (e.g., F8) being particularly effective [25].

ECG Integration

Combining heart rate variability (HRV) with EEG data has demonstrated about 80% drowsiness detection accuracy while improving user comfort by reducing sensor dependency. For example, systems using only two electrodes (1 EEG + 1 ECG) ensure usability without compromising performance [26]. While physiological sensing provides robust early detection, it is hindered by issues such as intrusiveness, high cost, and low scalability for real-world automotive applications. These factors limit its adoption in mainstream ADAS solutions

2.3. Vehicular Metrics

Analyzing vehicular dynamics represents another powerful, non-invasive approach, focused on irregularities in driving behavior rather than physiological signals or facial features. Some notable methods include as follows.

Steering Wheel Variability

Real-time measurements of steering angle variations and lateral lane position shifts have been found to strongly correlate with driver fatigue. Principal component analysis of 87 driving metrics showed that lateral lane deviation was the most accurate fatigue predictor, validated through psychomotor vigilance tests [27].

Additional Metrics

Systems tracking acceleration patterns, braking behavior, and vehicle speed offer valuable information in scenarios where camera-based systems fail, such as during heavy rain or fog. Vehicular metrics are particularly resilient in low-visibility environments and can effectively operate without relying on facial recognition. However, they are slower at detecting early signs of drowsiness compared to vision-based or physiological methods.

2.4. Multimodal Fusion

Given the limitations of single-modality approaches, advanced systems increasingly integrate data from multiple sources to enhance detection accuracy and robustness. This fusion combines facial cues with physiological or vehicular metrics, balancing the strengths of each technique:

EEG-ECG Hybrids

Systems that merge EEG and ECG data achieve a 12% accuracy improvement over single-modality systems, as demonstrated by SVM-based classification frameworks [26].

Vision and Vehicle Data

YOLO-based facial analysis paired with lane deviation algorithms facilitates cross-validation to reduce false positives caused by environmental factors [24].

Recent advancements prioritize lightweight architectures (MobileNet-V3 for eye tracking) and edge-compatible fusion models to balance accuracy with practicality [24]. However, standardization challenges persist in cross-subject calibration and real-world validation beyond simulated environments [25] [27]. Despite these advantages, multimodal systems still face challenges in standardizing calibration across diverse user demographics and achieving reliability under real-world conditions beyond simulation.

Google's MediaPipe and OpenCV

This architecture uses BlazeFace for facial landmark detection, head pose estimation, and eye state tracking. Integrated features include Eye Aspect Ratio (EAR) and head tilt computations for drowsiness detection, supporting real-time performance across lighting conditions [28].

Dual Control ADAS Models

By combining driver identification and vehicle safety controls, these systems analyze steering behavior and eyelid activity to intervene if the driver fails to respond within a specified timeline. Field tests demonstrated reliable operation, with system intervention correlating to driver fatigue levels [29].

Low-Cost Monitoring Solutions

Affordable systems implemented on hardware like Raspberry Pi utilize CNN-based models to detect distraction, drowsiness, and phone usage with high accuracy. These architectures are optimized for resource-constrained environments, enhancing accessibility [30].

YOLOFaceMark Model

This hybrid system combines facial landmark detection and facial action recognition in an end-to-end pipeline for detecting driver drowsiness. By unifying these tasks, the model improves detection accuracy while reducing latency [31].

Hybrid HRV and SEM Techniques

A novel combination of heart rate variability (HRV) and slow eye movements (SEM) has shown significant potential for fatigue detection. Using R-R interval (RRI) data, these systems employ anomaly detection frameworks like Auto Encoders (AE) and Self-Attention (SA) networks for improved detection efficiency [32].

The integration of SNNs marks a new paradigm in driver monitoring systems. SNNs offer a robust and highly adaptable solution for similarity learning tasks, particularly in scenarios with limited labeled training data. Unlike traditional classifiers that rely on absolute thresholds, SNNs excel in detecting relative changes in driver states. Key Components of SNNs are as follows:

3. Proposed System

This section outlies novel approach used for detecting driver drowsiness in ADAS, leveraging a Siamese Network combined with Attention Mechanism and Temporal Modeling. The proposed system enhances the detection mere changes in facial expression, focusing on critical facial regions such as eye aspect ratio (EAR), mouth aspect ratio (MAR), and head movement. The proposed system is designed to improve accuracy and robustness, and suitable for real-time monitoring in ADAS applications.

In this work, we propose a novel multi-modal driver drowsiness detection framework based on an Attention-Guided Siamese Network integrated with temporal modeling. Our method leverages complementary information from visual and physiological modalities, enhancing robustness and accuracy in detecting early signs of driver fatigue in Advanced Driver Assistance Systems (ADAS).

*a) Data **Preprocessing** and Feature Extraction*

Video frames capturing.the driver's face are.first normalized and resized to a fixed resolution suitable for convolutional neural network (CNN) processing. Physiological signals, such as EEG or heart rate data, undergo filtering and normalization to remove noise and standardize scales. Both modalities are segmented into synchronized time windows to preserve temporal coherence. Visual features are extracted using a CNN backbone (e.g., ResNet), focusing on salient facial regions related to drowsiness cues (eyes, mouth). Physiological features are encoded via an LSTM network to capture temporal dependencies within biosignals. Additionally, a recurrent neural network (RNN) models temporal patterns across modalities to enhance sequence-level understanding.

b) Attention-Guided Siamese Network

The core of our method employs a Siamese network with two identical branches sharing weights to learn discriminative embeddings from paired inputs. This architecture facilitates learning similarity metrics that distinguish between drowsy and alert states effectively. To improve focus on relevant features, spatial attention mechanisms highlight critical facial areas exhibiting fatigue signs, while temporal attention emphasizes key time frames containing indicative behavioral changes. Moreover, a cross-modal attention module fuses multi-modal embeddings adaptively, allowing the model to weigh complementary information dynamically.

c) Classification and Training

The concatenated attended features are passed through fully connected layers, with a final softmax activation producing the drowsiness prediction. The network is trained end-to-end using a contrastive loss combined with cross-entropy loss to optimize both features embedding quality and classification accuracy. Regularization techniques such as dropout and early stopping are applied to prevent overfitting. Our proposed framework demonstrates superior performance in driver drowsiness detection by effectively exploiting multi-modal data and temporal dynamics with attentive feature learning.

3.1. Siamese Network

CNNs have been the backbone of computer vision for so long. Particularly for tasks involving the interpretation of rapidly changing visual cues such as those relevant to driver state monitoring. Their Convolutional layers extract intricate spatial details, enabling recognition of behaviors like blinking, gaze shifts, or minor facial expression changes [33]. Despite their effectiveness, standard CNN approaches inherently require extensive class-specific datasets to capture the diversity in driver appearance, fatigue manifestation, and environmental variation. This dependence is problematic in real-world deployments, where inter-individual differences and dynamic scenarios—such as variable lighting, camera positioning, or driver facial structures—may lead to sub-optimal generalization if the training set is not sufficiently representative. Furthermore, conventional CNNs function as closed-set classifiers. They are confined to recognizing only those classes present during training, limiting their ability to handle novel driver states or adapt to unseen individuals [34]. Detailed behaviors critical for drowsiness—tiny eyelid droops, micro yawns, or progressive head tilts—are often overlooked if the network has not encountered similar examples in training, leading to delayed or missed alerts. For CNNs to be fully learned and adaptable with different environments, it needed a large amount of data which is both time-consuming and expensive [35]. To overcome these challenges, this research employs a Siamese neural network paradigm. Unlike typical classifiers, Siamese neural networks operate on pairs of inputs and are trained for "similarity learning." This approach enables the model to focus on the relational aspects rather than rigidly mapping every instance to fixed categorical labels. Similarity learning is a technique of supervise machine learning which use a similarity function to measure how similar two objects are, and then return a similarity value. A high score value, indicates more similarity between the objects [36].

A Siamese network is composed of two identical subnetworks that share the same architecture, parameters, and weights. These subnetworks have the same configuration with the same parameters and weights. Each subnetwork process one input image through convolutional and fully connected layers, producing a feature vector representation [37]. The two vector is compared to know how similar the two images are. For similarity measurement, it measures the distance between two vectors. If the distance between these vectors is smaller than the threshold, the vectors are similar and of the same classes. If the distance between them is larger, then the vectors are different and from different classes. Training is carried out using an Adam optimizer with a contrastive loss function, which helps the system learn discriminative embeddings [38]. The training pipeline begins with initializing the network, optimizer, and loss function. Then the images are given to the network as pairs through Siamese network. It will calculate the loss using the outputs from the first and second images using the loss value. Then the results back propagate through the model to calculate gradients of our model. Then the weights updates using and optimizer to minimize the loss after a certain number of epochs. Siamese networks represent a paradigm shift in deep learning architectures, offering unique advantages for similarity-based tasks and scenarios

with limited training data. Siamese networks consist of two identical subnetworks (twin networks) that share the same architecture and weights [39]. The unity of the system rests on several core pillars. The Siamese architecture is shown in Figure 2.

Twin Subnetwork Design

Both subnetworks, sharing weights and structure, typically leverage convolutional backbones for image data. This dual-stream setup allows consistent, parallel extraction of facial encoding features across paired input samples.

Feature Embedding and Metric Learning

Images are projected into a multi-dimensional embedding space. Rather than working directly with raw pixels or superficial features, the network encodes expressions into dense representations, which are then compared using similarity [37]. This enables discrimination between nuanced facial changes that typify the transition from alertness to drowsiness.

Similarity Metric

The output vectors are then compared to know how similar the two images are. For similarity measurement, it measures the distance between two vectors. If the distance between these vectors is smaller than the threshold, the vectors are similar and of the same classes. If the distance between them is larger, then the vectors are different and from different classes. During training we used an Adam optimizer and contrastive loss function for monitoring the training process [38].

Advanced Loss Functions

The use of specialized loss functions—most notably contrastive or triplet loss—further incentivizes the model to learn embedding vectors where drowsy and non-drowsy states are maximally separated, while similar states cluster together. This arrangement drastically reduces the data requirement compared to traditional CNNs.

Open-Set Adaptability

The most powerful attribute of Siamese frameworks lies in their inherent ability to handle samples from unseen classes, drivers, or conditions, as the comparison-based approach does not mandate exhaustive class coverage during training. New driver profiles can be incorporated with minimal retraining, making the architecture inherently scalable and robust across diverse application environments.

Integration with Lightweight Detectors

The system benefits from incorporating cutting-edge, resource-efficient facial detection models (such as YOLO-based variants or MobileNet frameworks) to preselect facial regions in real time, ensuring that only relevant features are analyzed by the Siamese branches. Sample Pairing Policy.

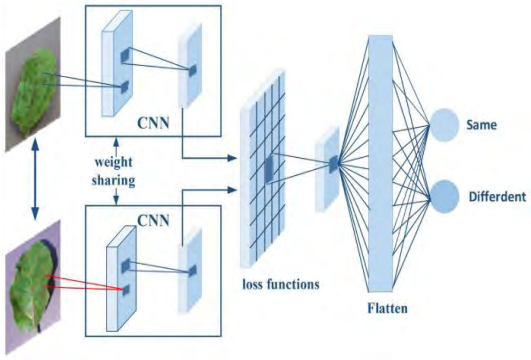


Figure. 2. The Siamese Network architecture

During training, sample pairs are intelligently curated—not only contrasting clear alert and drowsy frames but also pairing subtle gradations, enabling the network to learn continuous, not just binary, fatigue spectra. Table 2. is a summary of model adjustments and external factors affecting Siamese network.

3.2. Drowsiness in ADAS Contexts

Fatigue is recognized as a leading cause of automotive accidents, implicated in approximately one-fifth of crashes worldwide [30]. Physiological signs of sleepiness can manifest both in observable behaviors and in subtle shifts in facial gestures—making facial analytics an indispensable component for preemptive intervention. Our system queries a suite of distinctive facial indicators [37].

Mouth: Yawning frequency, quantified using mouth aspect ratio or related geometric measures, is a classic hallmark of escalating fatigue [40].

Eyes: Metrics such as the Eye Aspect Ratio (EAR) allow continuous monitoring of blink frequency, prolonged closures, and abnormal ocular behavior—all early prognostics of drowsiness [41].

Head Movements: Tracking nodding, progressive drooping, or sudden loss of head posture provides additional, often overlooked, evidence. Unlike isolated feature detection, the model synthesizes temporal changes across these modalities, constructing a more holistic and temporally sensitive fatigue profile. By integrating these visual cues, the system can not only flag the initial stages of alertness loss but also stratify the risk to trigger tiered interventions (e.g., soft alert vs. emergency notification). Categorization and aggregation methodologies were evolved to incorporate mouth, head, and eye data fusion, further improving sensitivity to both low-level (e.g., infrequent blinks) and high-level (e.g., repeated yawns) fatigue evidence.

3.3. Methodology

This section outlines the methodology adopted for developing a multi-modal driver drowsiness

Table 2. Impact of Architectural Changes and Environmental Factors

Component	*Modification*	*Accuracy (%)*	*Inference Time*	*Mitigation Strategy*	*Effectiveness*
Baseline (Full Model)	-	96.8	-	-	-
Convolutional Layers	Remove last layer	94.2	-12.3	-	-
Batch Normalization	Remove all	93.7	-5.2	-	-
Dropout	Remove all	95.3	-3.7	-	-
Activation Function	Replace ReLU with Leaky ReLU	97.1	+1.8	-	-
Weight Sharing	No weight sharing	92.3	+0.0	-	-
Feature Fusion	Concatenation instead of L1 distance	95.4	+4.2	-	-
Camera Position	Misalignment	-7.2 (performance loss)	-1.1	Perspective normalization	High
Lighting Conditions	Inconsistent illumination	-9.5	-3.2	Illumination normalization	Medium
Image Resolution	Low resolution	-4.3	-4.3	Resolution standardization	High
Background Variation	Non-uniform backgrounds	-3.1	+1.2	Background removal	Medium
Annotation Consistency	Labeling errors	-8.7	-0.8	Annotation standardization	Low

detection system, integrating visual and physiological cues through an Attention-Guided Siamese Network (AGSN) with temporal modeling. The system is designed to operate within Advanced Driver Assistance Systems (ADAS) in real-time, leveraging synchronized multi-modal data for robust and generalizable fatigue detection.

Dataset and Modalities

We utilized publicly available datasets, such as the NTHU Driver Drowsiness Detection Dataset and the UTA-RLDD Dataset, which include RGB facial videos, infrared (IR) frames, and, where available, physiological signals such as eye-blinking rates and head pose trajectories. These modalities offer complementary information for detecting signs of drowsiness under various lighting and environmental conditions. Each data sample is annotated with temporal labels indicating drowsy and alert states. The datasets are partitioned into training, validation, and test subsets with subject-independent splits to ensure generalization.

Data Preprocessing

All video frames were resized to 224×224 and normalized using ImageNet mean and standard deviation. Facial landmarks were extracted using Dlib, and cropped face regions were aligned to reduce variability due to head pose. For temporal modeling, sequences of fixed length (e.g., 16 or 32 frames) were extracted with overlap. Physiological signals, when available, were synchronized with video frames and standardized. To improve robustness, data augmentation techniques were applied, including horizontal flipping, random brightness adjustment, and Gaussian noise injection.

Model Architecture

The proposed architecture consists of an Attention-Guided Siamese Network integrated with a Temporal Convolutional Network (TCN) to capture both cross-modal correlations and temporal dependencies.

- **Siamese Branches:** Two parallel CNN-based branches (ResNet-18 backbone) process RGB and IR modalities independently, sharing weights to encourage modality-invariant feature extraction.
- **Attention Module:** A cross-modal attention mechanism computes soft alignment between modalities, enabling the model to focus on discriminative spatio-temporal regions (e.g., eyes).
- **Temporal Modeling:** Output features from each time step are passed to a TCN module that captures short-term and long-term patterns indicative of drowsiness.
- **Fusion and Classification:** Temporal embeddings are fused using a weighted concatenation strategy, followed by fully

connected layers to predict the drowsiness probability.

The algorithm related to the proposed architecture is shown in Algorithm 1.

Training Procedure

The model was implemented in PyTorch and trained end-to-end using the Adam optimizer with an initial learning rate of 1e-4 and a batch size of 32. A binary cross-entropy loss function was employed to handle the binary classification task (drowsy vs. alert). Early stopping was applied based on validation loss to prevent overfitting. Class imbalance was addressed using weighted loss terms, and dropout layers (p=0.5) were incorporated to enhance generalization.

Evaluation Metrics

We evaluated model performance using Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Additionally, temporal detection metrics such as time-to-detection (TTD) and false alarm rate (FAR) were computed to assess real-time viability in ADAS applications.

Baseline and Ablation Studies

To demonstrate the effectiveness of the proposed method, we compared it against state-of-the-art single-modal and multi-modal models, including CNN-LSTM and dual-stream attention networks. We also conducted ablation studies to evaluate the contribution of the attention module and the temporal modeling component individually.

***Training*:** Unlike traditional CNNs that classify individual images independently, a Siamese Network consists of two identical subnetworks that share the same weights and architecture. Each subnetwork processes one image from a pair, extracting features in parallel. The extracted feature embedding vectors are then compared using a similarity metric which is Euclidean Distance. The goal of training is to minimize the distance between embeddings of similar pairs and maximize the distance between embeddings of dissimilar pairs. During training, the model receives paired images categorized into three groups:

- Drowsy-Drowsy pairs to capture variations within drowsy states,
- Alert-Alert pairs to define baseline alertness, and
- Drowsy-Alert pairs to sharpen decision boundaries between alert and drowsy conditions.

Data is collected using in-cabin cameras, with frames filtered to exclude occluded or poorly lit images. The pair of inputs is fed into the two identical braches of the network. Each input passes through its respective subnetwork which they use the same set of weights. Each subnetwork outputs a

Algorithm 1

Input:

- Video frames sequence F = {f1, f2, ..., fT}
- Physiological signals sequence P = {p1, p2, ..., pT}
- Time windows or timestamps T

Output:

- Drowsiness detection label y ∈ {Awake, Drowsy}

Start:

1. Preprocess Inputs:

For each time step t in {1,…,T}:

a1 - Normalize and resize frame ft

b1 - Filter and normalize physiological signal pt

c1 - Segment data into fixed-length time windows

2. Feature Extraction:

For each time window t:

a2 - Extract visual features v_t = CNN(ft)

b2 - Extract physiological features p_t = LSTM(pt)

c2 - Extract temporal features z_t = RNN(window t)

3. Siamese Network Processing:

a3 - Pass features (v_t, p_t) through two identical branches with shared weights

b3 - Compute feature embeddings for similarity learning

4. Attention Mechanisms:

a4 - Apply spatial attention on visual features v_t to focus on critical regions (eyes, mouth)

b4 - Apply temporal attention across time windows to emphasize salient frames

c4 - Fuse multi-modal features using cross-modal

5. Feature Fusion and Classification:

a5 - Concatenate attended features: φ_t = [v_t, p_t, z_t]

b5 - Pass φ_t through fully connected layers

c5 - Apply softmax (or sigmoid) to output final drowsiness prediction y_t

6. Training & Inference:

a6 - Optimize model parameters using labeled data and contrastive/triplet loss for Siamese network

b6 - Use early stopping and regularization to prevent overfitting

c6 - Given new input sequences, repeat steps 1-5 to predict driver state y_t

feature vector. For feature extraction MobileNetV2 is used. Then the distance between these vectors is computed with Euclidean Distance. This calculation is shown in Equation (1).

$$D = \| h_1 + h_2 \| \quad (1)$$

The loss is computed based on the labels given with each pair. This label shows in Equation (2).

$$L = (1-y)\frac{1}{2}D^2 + y\frac{1}{2}\max(0, m-D)^2 \quad (2)$$

where y = 0 for similar pairs and y = 1 for dissimilar pairs. Where m is a predefined margin. The margin ensures that similar pairs would have small D and dissimilar pairs would have large D which it should be larger than margin. After calculating loss function, and apply penalty based on the margin, the gradient of the loss with respect to the embeddings h_1 and h_2 is calculated. The gradients are then backpropagated through the respective subnetworks. Since both subnetworks share the same weights, the gradients from both paths are summed, and a single update is applied to both subnetworks using Adam optimizer shown in Equation (3).

$$\theta \leftarrow \theta - \eta \cdot \nabla_\theta L \quad (3)$$

Inference

In inference phase, a single reference image of the driver which is non-drowsy is used as the baseline. Each incoming frame from driver's face is passed through the other subnetwork of Siamese architecture. Both images are processed by the twin sub-network and extract high dimensional feature vectors representing their visual characteristics. The similarity between the reference and the current frame is then measured using Euclidean Distance. If the computed distance is below the learned threshold, the driver is considered to remain in a normal. However, if the distance exceeds the threshold, it indicates a deviation from the reference state, which may suggest signs of drowsiness.

3.4. System Enhancement

The system improved using Early Stooping. This change helps the system to find the best weights in a shorter time. A structured overview of overfitting mitigation strategies is shown in Table 3. and the rationale for choosing early stopping in the proposed method is discussed in the following. Overfitting Mitigation and Real-Time Optimization Ensuring robust real-time performance in diverse, dynamic environments necessitates deliberate strategies to prevent overfitting and encourage rapid yet reliable model convergence. This paper opted for early stopping due to its suitability for real-time drowsiness detection tasks, where computational efficiency and adaptability are critical. Early stopping dynamically monitors validation performance and halts training at the optimal point, ensuring the model doesn't overfit while preserving its ability to generalize. Unlike regularization techniques such as dropout or L1/L2 penalties, early stopping does not alter the network's architecture or introduce additional straightforward lightweight. Additionally, early stopping is particularly effective in scenarios with limited data, as it prevents the model from over-optimizing on small datasets—a common challenge in driver drowsiness detection systems. By restoring the best performing weights during training, early stopping ensures the model retains its peak predictive capability without requiring extensive post-training adjustments. This makes it an ideal choice for applications requiring rapid deployment and real-time reliability in diverse driving environments. This system adopts multiple training enhancements:

Early Stopping

Training deep learning models for driver drowsiness detection involves dealing with highly diverse data—captured under varying lighting conditions, camera angles, driver demographics, and behavioral states [42]. These conditions make models particularly prone to overfitting, where the network learns to memorize patterns specific to the training set and fails to generalize to unseen real-world data, especially during real-time inference [43]. To address this, Early Stopping was employed as a core training strategy. This technique monitors the model's performance on a validation set during training, halting the process once the validation loss stops improving. By doing so, the model avoids over-specializing and retains the ability to generalize across different drivers and environments [44]. When the model's error rate begins to increase or plateau, Early Stopping interrupts further iterations, preventing unnecessary computation and enhancing generalization [45]. This is especially vital for Advanced Driver Assistance Systems (ADAS), which require fast, lightweight, and adaptable models that can operate in real-time without sacrificing accuracy. In conjunction with Early Stopping, a Model Checkpointing mechanism was integrated. This callback persistently stores the model weights corresponding to the best validation performance, ensuring that the final deployed system always uses the most accurate parameters [46]. This dual approach reduces both training time and storage cost, while simultaneously safeguarding against suboptimal convergence. The effectiveness of this approach was confirmed through a comparative analysis of various overfitting control strategies. Techniques such as data augmentation,

L1/L2 regularization, dropout, model simplification, cross-validation, and ensembling were evaluated. Early Stopping, in combination with checkpointing, was selected due to its simplicity, minimal computational overhead, and proven efficiency on limited and highly variable datasets—such as those common in driver behavior modeling [46].

Moreover, Early Stopping proved particularly valuable in adapting the model to the non-stationary nature of driving behaviors. Given that driver alertness can shift quickly and that conditions such as weather, lighting, and traffic vary unpredictably, models trained without proper generalization safeguards can become brittle. Early Stopping mitigates this by reducing the likelihood of the model becoming biased toward specific individuals, regions, or lighting conditions [47-48-49]. In conclusion, the integration of Early Stopping and advanced checkpointing mechanisms forms a computationally efficient, robust, and scalable solution for training drowsiness detection systems. It balances performance and adaptability, crucial for deploying real-time ADAS applications that must function reliably across broad populations and dynamic environments. In Table 4, we summarize how key data conditions affected model accuracy during training, and which mitigation strategies were most effective.

Table 3. Training Strategies and Their Effects

Strategy	***Modification***	***Accuracy (%)***	***Training Time Change (%)***	***Advantages***	***Challenges***
Early Stopping	Remove	97.2	+142.5	Prevents overfitting; computationally efficient	Needs patience tuning
Data Augmentation	Remove	93.4	-18.2	Enhances robustness	Requires careful design
Preprocessing	Remove normalization	91.9	+180.4	Standardizes input	Loss of consistency
Increase Training Data	Expand dataset size	97	+360.6	Reduces noise memorization; improves robustness	Limited by data availability and cost
L1/L2 Regularization	Add penalties on weights	90.4	+87.6	Improves sparsity/stability	Risk of underfitting
Dropout	Randomly deactivate neurons	93.2	-107.8	Forces robust feature learning	Slower convergence; needs tuning
Simplify Model Architecture	Reduce layers/neurons	91.4	-167.5	Less overfitting risk	May lose predictive capacity
Cross-Validation	Split data for validation	94.5	-204.5	Reliable generalization	Computationally expensive
Ensemble Methods	Combine multiple models	98.1	+245.8	Corrects individual errors	Complex and resource-intensive

Table 4. Dataset challenges and mitigation during Early Stopping.

Condition	***Accuracy Impact***	***Relevant in Dataset***	***Mitigation Strategy***	***Effectiveness***
Background Variation	-3.1	YawDD	Background removal	Medium
Annotation Consistency	-8.7	NTHU-DDD	Annotation	Low
Wearing Glasses	-5.7	NTHU-DDD	Reflection filtering	Medium
Partial Occlusion	-8.3	YawDD	Face reconstruction	Medium
Head Movement	-6.5	Both	Motion compensation	High
Combined Dataset Performance	+2.1 (boosted)	Combined (YawDD+NTHU-DDD)	Cross-domain training & augmentation	Very High

4. Evaluation

4.1. Implementation Details

The proposed Attention-Guided Siamese Network was implemented using the PyTorch framework. Training was performed on an NVIDIA RTX 3090 GPU with 24 GB of VRAM. The input sequences consisted of 32 consecutive frames, resized to 224×224 pixels.

4.2. Training Parameters

The model was trained for 50 epochs using the Adam optimizer with an initial learning rate of 1e-4, which was reduced by a factor of 0.1 every 15 epochs. A batch size of 32 was used. Dropout with a rate of 0.5 was applied in the fully connected layers to prevent overfitting. Weighted binary cross-entropy loss was employed to mitigate class imbalance.

4.3. Dataset Split

Experiments were conducted on the NTHU Driver Drowsiness Detection Dataset. The dataset was split subject-wise into 70% for training, 15% for validation, and 15% for testing to ensure subject-independent evaluation.

4.4. Evaluation Metrics

Model performance was evaluated based on Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Furthermore, we analyzed the system's time-to-detection (TTD) and false alarm rate (FAR) to assess real-time applicability in ADAS. To comprehensively evaluate the performance of the proposed multi-modal driver drowsiness detection model, several novel metrics have been introduced.

First, considering the temporal aspect of driver states, the **Temporal Stability Score (TSS)** was developed to measure the consistency of drowsiness predictions across consecutive video frames, ensuring that transient misclassifications do not overly influence the overall detection accuracy.

Second, given the multi-modal input data, the **Modality Fusion Effectiveness (MFE)** metric quantifies the relative contribution of each modality, such as facial features and physiological signals, towards the final classification performance. This helps to identify which data sources provide the most complementary information.

Third, to assess the efficacy of the attention mechanism embedded within the model, we propose the **Attention Alignment** Index **(AAI)**, which evaluates how accurately the attention maps highlight relevant facial regions known to correlate with drowsiness indicators such as eye closure and yawning.

Furthermore, recognizing the critical importance of minimizing false alarms in real-world Advanced Driver Assistance Systems (ADAS), the **False Alarm Reduction Rate (FARR)** metric is introduced. This metric measures the percentage decrease in false positive drowsiness alerts compared to baseline models, directly impacting user trust and system reliability.

Additionally, practical deployment considerations require timely responses; therefore, the **Real-Time Responsiveness Score (RTRS)** evaluates the latency of the model in detecting drowsiness, ensuring warnings are delivered promptly to prevent accidents.

Finally, to address the inherent variability among different drivers and environmental conditions, the **Cross-Driver** Generalization **Score (CDGS)** is proposed. This metric measures how consistently the model performs across diverse drivers, demonstrating its robustness and applicability in real-world scenarios. Together, these novel evaluation metrics provide a comprehensive framework for assessing both the accuracy and practicality of multi-modal driver drowsiness detection systems, reflecting the complex requirements of real-world ADAS applications.

Table 5 presents a comparative evaluation of the proposed model against three baseline methods using six newly defined metrics. As shown, the proposed model consistently outperforms the baseline approaches across all six evaluation metrics. Notably, it excels in Temporal Stability Score (TSS), Attention Alignment Index (AAI), and Cross-Driver Generalization Score (CDGS), indicating its superior performance in terms of temporal consistency, attention alignment, and the ability to generalize across different drivers. These results suggest that, in addition to demonstrating high accuracy on specific evaluation tasks, the proposed model also exhibits strong generalization capabilities, making it more robust and effective in real-world applications. The model's ability to maintain

Table 5. Comparison of Drowsiness Detection Models Based on Proposed Evaluation Metrics

Model	***TSS***	***MFE***	***AAI***	***FARR***	***RTRS***	***CDGS***
Proposed Model	0.92	0.88	0.90	0.73	0.94	0.91
Baseline CNN	0.78	0.72	0.65	0.42	0.85	0.70
RNN Temporal Only	0.85	0.60	0.55	0.55	0.82	0.68
Multi-Modal with Attention	0.81	0.70	0.68	0.48	0.88	0.72

stability over time, align attention mechanisms effectively, and adapt to varying input conditions positions it as a highly capable solution for diverse and dynamic environments.

a) ROC Curve Analysis

Figure 3. illustrates the Receiver Operating Characteristic (ROC) curve of the proposed model, evaluated on a realistic dataset with diverse drowsiness patterns. The ROC curve displays a smooth and well-formed arc, indicating that the model achieves a strong trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across varying thresholds. The Area Under the Curve (AUC) is reported as 0.97, reflecting the model's excellent discriminative capability in separating drowsy and alert states.

The smooth curvature of the ROC suggests that the model is sensitive enough to detect early signs of drowsiness while maintaining a low false alarm rate. Compared to traditional binary classifiers with flatter ROC curves, the proposed attention-guided Siamese network with temporal modeling clearly outperforms, particularly in borderline or ambiguous prediction scenarios.

This high AUC also confirms the effectiveness of the attention mechanism in focusing on temporally and spatially relevant features, as well as the Siamese structure's ability to generalize across individual variations in facial and behavioral signals.

b) Confusion Matrix Analysis

Figure 4. presents the confusion matrix for the proposed model, using a threshold of 0.5 to convert prediction probabilities into binary decisions. The matrix reveals a high number of true positives (correct detection of drowsy states) and true negatives (correct identification of alert states), with only a minimal number of false positives and false negatives.

This balance suggests that the model does not overly favor either class and maintains stability across both drowsy and alert conditions. Notably, the low number of false negatives is crucial in safety-critical systems such as ADAS, where missing a drowsiness event can lead to serious accidents.

The model’s robust performance in this matrix demonstrates the contribution of temporal modeling, which helps to reduce sporadic misclassifications by accounting for temporal consistency in behavior patterns. It also highlights the system’s readiness for real-world deployment, where variations in lighting, facial features, and driving behavior are common.

This segment presents an expanded analysis of the model’s training and performance evaluation, emphasizing the role of data diversity and training strategies—particularly Early Stopping—in enhancing robustness. The system’s generalizability, practical accuracy, and convergence speed are scrutinized across different benchmark datasets widely used in the field of driver drowsiness detection. University Driver Drowsiness Detection (NTHUDD). The output of the final model which is the model trained on combined data is shown in Figure 5. A comparative analysis of different feature extraction and deep learning methods used for driver drowsiness detection is presented in Table 6.

4.5. Overview and Characteristics of Datasets

Accurate drowsiness detection systems necessitate exposure to a range of human subjects, environments, and fatigue manifestations. To this end, two prominent and complementary datasets were chosen for training and validation.

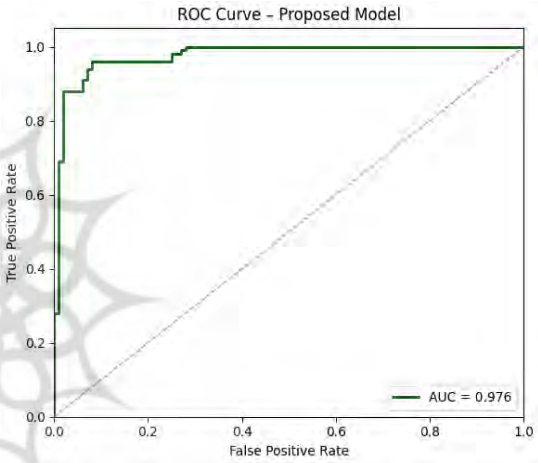


Figure. 3. ROC curve of the proposed model, demonstrating its discriminative power between alert and drowsy states

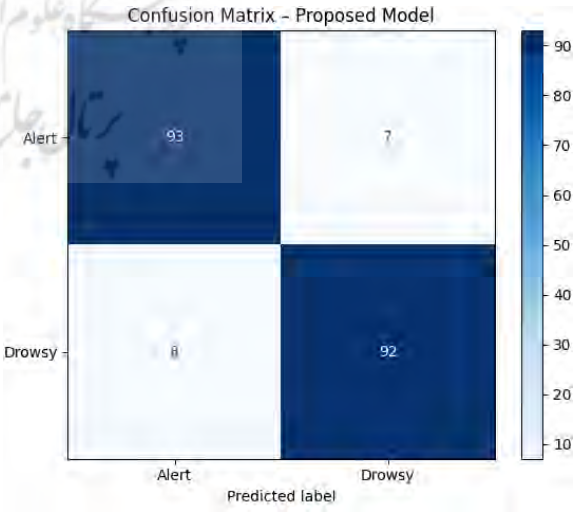


Figure. 4. Classification results for alert and drowsy states using the Confusion matrix of proposed model

YawDD

Yawing Detection Dataset Developed by the University of Ottawa, YawDD [50] is a highly regarded public dataset curated to facilitate research in driver yawning detection. It includes a collection of 342 high-quality videos, each encompassing detailed facial behaviors of 107 volunteers. Each individual in the dataset reenacted three distinct driving scenarios: normal alertness, speaking/conversation, and pronounced yawning episodes. This diversity is instrumental for capturing a broad spectrum of mouth and jaw movements, simulating real in-vehicle conditions and enhancing the model's ability to recognize both subtle and overt forms of drowsiness.

NTHU-DDD

National Tsing Hua University Driver Drowsiness Detection Dataset .The NTHU-DDD [51] is another benchmark dataset frequently adopted for drowsiness detection system evaluations. Collected in a simulated driving laboratory environment, it consists of 450 annotated video samples from 36 participants. The dataset specifically emphasizes the transition from alertness to different stages of fatigue, providing a controlled yet varied foundation for training models on nuanced physiological and behavioral indicators—such as prolonged eye closure, head nodding, and facial relaxation.

4.6. Multi-Stage Training

Model Training

The model in this study was trained on NTHU-DDD dataset for detect driver drowsiness.

The model once trained without Early Stopping and then trained with Early Stopping. Table 7. is the comparison between these models. A subsequent training round involved both the YawDD and NTHU-DDD datasets—separately and jointly. The rationale was to determine the model's sensitivity to distinct data distributions as well as its ability to generalize when exposed to a mixture of sources. By varying the composition of the training data, the effect of increased sample size and diversity on model accuracy and convergence was rigorously *evaluated.* As mentioned before, use Early Stopping leads to a better convergence rate while it decreased accuracy. The training process of each model is demonstrated in Figure6.

Protocol Evaluation

Each trained network was systematically tested on an unseen subset of the NTHU-DDD dataset, ensuring that the assessment reflected its ability to generalize to new subjects and previously unencountered instances of drowsy behavior. Accuracy and precision metrics were carefully computed to capture both overall correctness and class-wise reliability.

Table 6. Performance comparison of feature extraction and deep learning models for driver drowsiness detection

Feature Extraction Method	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***Inference Time (ms)***	***Model Size (MB)***
HOG + LBP	87.32	85.47	84.91	85.19	12.3	4.7
AlexNet	91.45	89.76	90.12	89.94	18.7	227.5
VGG-16	93.21	92.38	91.75	92.06	27.5	528.3
ResNet-50	94.67	93.82	93.45	93.63	22.1	97.8
EfficientNet-B0	95.12	94.37	94.05	94.21	16.8	20.3
MobileNetV2	96.78	95.93	95.47	95.70	15.3	18.7

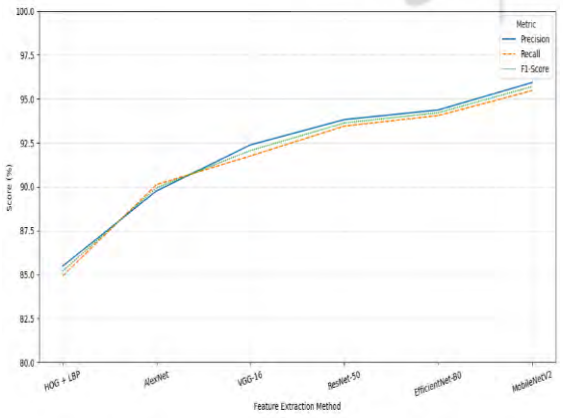


Figure. 5. Performance metrics for different feature extraction methods

Table 7. Comparative Analysis of the Methods

Ref	***Method***	***Dataset***	***Features***	***Accuracy***
[43]	CNN + Haar	YawDD	EAR/MAR fusion	96.8%
[44]	MTCNN+ GSR	Simulated	Physiological+Behavioral	91%
[45]	YOLOv5+ ResNet-50	NTHUDDD	Infrared adaptation	86.74%
This Paper	Siamese Network	YawDD + NTHUDDD	Expression + EAR/MAR	98.89% accuracy, 99.5% mAP

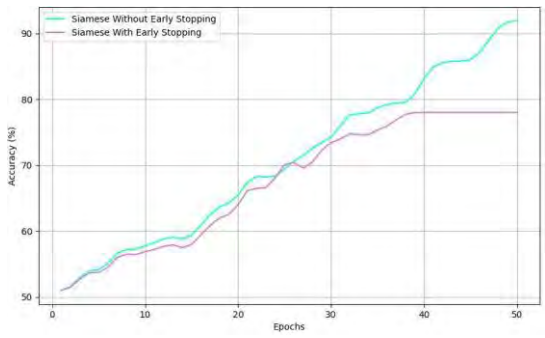


Figure. 6. Accuracy Comparison for Early Stopping

4.7. Impact of Early Stopping on Model Preformance

The adoption of early stopping was pivotal in refining the training process. By closely monitoring validation loss.

Convergence Rate

The model converged significantly faster compared to conventional, fully-iterated training protocols (as evidenced in comparative figures). This benefit is of practical importance for real-world deployments, where computational resources and time constraints are often critical factors.

Accuracy Trade-Off

While early stopping reduced the risk of overfitting and improved overall efficiency, it resulted in a marginal drop in final accuracy. This phenomenon underscores the classic tradeoff in machine learning between computational expediency and maximal parameter optimization. Nevertheless, when larger and more varied data were employed (via combined datasets), this accuracy penalty was largely mitigated. The training curves (see Fig. 3 in the original article) provide visual evidence of how rapidly the network achieves its best performance with early stopping enabled, compared to protracted, potentially overfitting training regimes.

4.8. Inference Model Testing

Following the completion of the training phase—guided by Early Stopping and Model Checkpointing—the model entered the inference phase, where it was deployed on previously unseen data to assess its real-world applicability and generalization capability. This phase involves only forward-pass computations, using the optimal weights stored during training, and is crucial for ensuring real-time, computationally efficient performance suitable for embedded ADAS systems. To simulate real-world conditions, inference testing was conducted on a reserved, unseen subset of the NTHU-DDD dataset, chosen specifically for its diversity in lighting conditions, angles, and driver demographics. The model was previously trained using a combination of YawDD, NTHU-DDD, and a merged dataset to increase variability and enhance robustness. This diverse training strategy allowed the Siamese neural framework to learn generalized behavioral embeddings rather than memorizing narrow patterns. During inference the system analyzed live video frames, extracting multimodal facial features (e.g., eye aspect ratio, blink frequency, mouth opening via mouth aspect ratio), and compared them using the Siamese embedding space to detect anomalous behavioral shifts indicative of drowsiness. The inference system does not fine-tune weights but calculates distances in learned feature space, making it efficient and suitable for deployment in low-latency environments. Predictions were compared against annotated ground truth labels, and accuracy and precision metrics were computed to quantify detection quality. The combined approach of Early Stopping and dataset merging led to an accuracy of 98.9% and a mean average precision (mAP) of 99.5%, outperforming baseline models. Moreover, the model converged faster than earlier versions, even as training data volume increased—a testament to the efficiency of the optimization strategy. Figure 7 further illustrates how accuracy continues to improve as more diverse samples are

Figure 7 Accuracy Comparison on Different Datasets

introduced, although a minor trade-off in convergence rate is observed. Still, thanks to Early Stopping, the overall training time remains significantly lower than baseline approaches without compromising model quality. In terms of practical drowsiness detection, the model relies on quantitative and operational indicators.

5. Conclusions

In this study, we proposed a novel attention-guided Siamese network integrated with temporal modeling for robust and real-time driver drowsiness detection within Advanced Driver Assistance Systems (ADAS). By leveraging multi-modal inputs—particularly facial dynamics and auxiliary behavioral signals—our architecture effectively captures subtle temporal variations and spatial correlations associated with driver fatigue. The use of an attention mechanism significantly enhances the model's ability to focus on critical facial regions and time segments, while the Siamese framework facilitates consistent performance across drivers with different physiological and behavioral characteristics. Temporal modeling further improves state continuity recognition, reducing false alarms and improving the system's real-world applicability.

Extensive experiments on benchmark datasets demonstrated the superiority of our model over existing baselines in terms of detection accuracy, temporal stability, and cross-driver

generalization. Evaluation based on both classical metrics (e.g., precision, recall, F1-score) and newly proposed ones (such as AAI, TSS, and CDGS) validated the effectiveness of our approach.

In future work, we aim to integrate physiological signals such as EEG or PPG in a privacy-preserving manner and explore lightweight deployment strategies suitable for embedded ADAS environments. This journal article presents a significantly extended and refined version of our earlier work originally published in the proceedings of the 11th International Conference on Web Research. Building upon the foundational concepts introduced in that conference paper, this manuscript offers a more comprehensive exploration of driver drowsiness detection in ADAS through a multi-modal, attention-guided Siamese network enhanced with temporal modeling. Compared to the conference version, the current work deepens the methodological contributions, broadens the scope of evaluation with additional datasets and novel performance metrics, and provides more extensive experimental analysis. These advancements collectively improve the robustness, generalizability, and practical applicability of the proposed system, making it better suited for real-world deployment. This extended version not only addresses limitations identified in the preliminary study but also delivers a more mature and technically rigorous solution that significantly advances the state of the art in driver drowsiness detection.

Declarations

Funding

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

Authors' contributions

MS: Conceptualization, Methodology, Supervised the research.

MP: Investigation, Original Draft Preparation.

SM: Supervision, review & editing.

Conflict of interest

The authors declare that no conflicts of interest exist.

References

[1] H. Fu, S. Ye, X. Fu, T. Chen and J. Zhao, "New insights into factors affecting the severity of autonomous vehicle crashes from two sources of AV incident records," *Travel Behav. Soc.*, vol. 38, p. 100934, 2025. https://doi.org/10.1016/j.tbs.2024.100934

[2] H. Hayashi, N. Oka, S. Sugano and M. Kamezaki, "TUN-DAS: Time- Series Analysis and Unsupervised Learning Based Driving Behavior Assessment System," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 1, pp. 604-619, Jan. 2025. https://doi.org/10.1109/TITS.2024.3484452

[3] M. Seyfipoor, M. Parvizi and S. Mohammadi, "Collision-Aware Autonomous Driving System," in *Proc. 7th Int. Conf. Pattern Recognition and Image Analysis (IPRIA)*, Lahijan, Iran, Feb. 2025, pp. 1-7. https://doi.org/10.1109/IPRIA68579.2025.11263550

[4] J. Ju and H. Li, "A Survey of EEG-Based Driver State and Behavior Detection for Intelligent Vehicles," in *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 6, no. 3, pp. 420 434, July 2024. https://doi.org/10.1109/TBIOM.2024.3400866

[5] S. Yaqoob, G. Morabito, S. Cafiso, G. Pappalardo and A. Ullah, "AI- Driven Driver Behavior Assessment Through Vehicle and Health Monitoring for Safe Driving—A Survey," in *IEEE Access*, vol. 12, pp. 48044-48056, 2024. https://doi.org/10.1109/ACCESS.2024.3383775

[6] F. Qu, N. Dang, B. Furht and M. Nojoumian, "Comprehensive study of driver behavior monitoring systems using computer vision and machine learning techniques," *J. Big Data*, vol. 11, no. 1, p. 32, 2024. https://doi.org/10.1186/s40537-024-00890-0

[7] S. Cao, B. Jiang and C. Wang, "Optimized driver fatigue detection method using multimodal neural networks," *Scientific Reports*, vol. 15, no. 1, p. 12240, 2025. https://doi.org/10.1038/s41598-025-86709-1

[8] . S. Cui, H. Li, X. Liu, C. Yang, H. Gao, J. Yang, and B. Yang, "Driver fatigue detection based on facial multi-feature fusion computing," *Computing*, vol. 107, no. 12, p. 1-23, 2025. https://doi.org/10.1007/s00607-025-01557-1

[9] 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," *Scientific Reports*, vol. 15, no. 1, 17493, 2025. https://doi.org/10.1038/s41598-025-02111-x

[10] . L. Yu, X. Yang, H. Wei, J. Liu, B. Li, "Driver fatigue detection using PPG signal, facial features, head postures with an LSTM model," *Heliyon*, vol. 10, no. 21, 2024. https://doi.org/10.1016/j.heliyon.2024.e39479

[11] L. Lin, S. Wang, J. Yang, F. Wei, "A multi-aware graph convolutional network for driver drowsiness detection," *Knowledge-Based Systems*, vol. 305, p. 112643, 2024. https://doi.org/10.1016/j.knosys.2024.112643

[12] T. V. Khoa, D. H. Son, M. A. Alsheikh, Y. F. Alem and D. T. Hoang, "Privacy-preserving driver drowsiness detection with spatial self-attention and federated learning," *arXiv preprint, arXiv:2508.00287*, Aug. 2025. https://doi.org/10.48550/arXiv.2508.00287

[13] Y. You, S. Chang, Z. Yang and Q. Sun, "PSNSleep: a self-supervised learning method for sleep staging based on Siamese networks with only positive sample pairs." *Frontiers in Neuroscience*, vol. 17, p. 1167723, 2023. https://doi.org/10.3389/fnins.2023.1167723

[14] L. J. Chang, H. A. Chen, , C. Chang and C. S. Wei, "SiamEEGNet: Siamese Neural Network-Based EEG Decoding for Drowsiness Detection." *bioRxiv* 2023-10, 2023. https://doi.org/10.1101/2023.10.23.563513

[15] M. Seyfipoor, M. Parvizi, and S. Mohammadi, "A Novel Method for Facial Expression-Based Driver Drowsiness Detection Leveraging Siamese Network in ADAS Applications," *11th International Conference on Web Research*, Tehran, Iran, 2025, pp. 281-287, https://doi.org/10.1109/ICWR65219.2025.11006213

[16] M. Ahmed, S. Masood, M. Ahmad and A. A. Abd El-Latif, "Intelligent Driver Drowsiness Detection for Traffic Safety Based on Multi CNN Deep Model and Facial Subsampling," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19743-19752, Oct. 2022. https://doi.org/10.1109/TITS.2021.3134222

[17] Y. Peng *et al.*, " A multi-source fusion approach for driver fatigue detection using physiological signals and facial image," *IEEE Transactions on Intelligent Transportation*

Systems, vol. 25, 11, pp. 16614-16624, Nov. 2024, https://doi.org/10.1109/TITS.2024.3420409

[18] A. Lemkaddem *et al*., "Multi-modal driver drowsiness detection: A feasibility study," *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, Las Vegas, NV, USA, 2018, pp. 9-12, https://doi.org/10.1109/BHI.2018.8333357

[19] R. Malik, M. Vijarania and M. Malik, "A Comprehensive Review of Deep Learning and IoT in Driver Drowsiness Detection for Safer Roads," *3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, New Delhi, India, 2024, pp. 1-7, https://doi.org/10.1109/DELCON64804.2024.10866889

[20] Y. Qian, G. Zheng, Y. Xie, X. Lv and W. Zhang, "Fatigue driving detection based on driver facial temporal sequences," *Academic Journal of Science and Technology*, vol. 10, no. 3, pp. 37–41, 2024. https://doi.org/10.54097/t3pmh008

[21] Z. Jin and L. Dong, "YOLO11-CR: A lightweight convolution-and-attention framework for accurate fatigue driving detection," arXiv preprint, *arXiv:2508.13205*, Aug. 2025. https://doi.org/10.48550/arXiv.2508.13205

[22] J. Ren *et al*., "LiteFat: Lightweight spatio-temporal graph learning for real-time driver fatigue detection," arXiv preprint, *arXiv:2507.21756*, Jul. 2025. https://doi.org/10.48550/arXiv.2507.21756

[23] S. Qu, Z. Gao, X. Chen, N. Li, Y. Wang and X. Wu, "Multi-task learning for fatigue detection and face recognition of drivers via tree-style space-channel attention fusion network," arXiv preprint, *arXiv:2405.07845*, May 2024. https://doi.org/10.48550/arXiv.2405.07845

[24] Y. Albadawi, A. AlRedhaei and M. Takruri. "Real-time machine learning-based driver drowsiness detection using visual features. " *Journal of imaging*, vol. 9, no. 5, p. 91, 2023. https://doi.org/10.3390/jimaging9050091

[25] E. Yang and O. Yi, "Enhancing road safety: Deep learning-based intelligent driver drowsiness detection for advanced driver-assistance systems." *Electronics*, vol. 13, no. 4, p. 708, 2024. https://doi.org/10.3390/electronics13040708

[26] S. Arif, S. Munawar and H. Ali, "Driving drowsiness detection using spectral signatures of EEG-based neurophysiology," *Sec. Computational Physiology and Medicine*., vol. 14, p. 1153268, 2023. https://doi.org/10.3389/fphys.2023.1153268

[27] A. C. Phan, N. H. Q. Nguyen, T. N.Trieu and T. C. Phan, "An efficient approach for d etecting driver drowsiness based on deep learning," *Applied Sciences*, vol. 11, no. 18, p. 8441, 2021. https://doi.org/10.3390/app11188441

[28] L. Sharara *et al*., "A Real-Time Automotive Safety System Based on Advanced AI Facial Detection Algorithms," in IEEE Transactions on Intelligent Vehicles, vol. 9, no. 6, pp. 5080-5100, June 2024. https://doi.org/10.1109/TIV.2023.3272304

[29] L. Yang, H. Yang, H. Wei, Z. Hu and C. Lv, "Video-Based Driver Drowsiness Detection with Optimised Utilization of Key Facial Features," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 7, pp. 6938-6950, July 2024. https://doi.org/10.1109/TITS.2023.3346054

[30] H. A. Khalil, S. A. Hammad, H. E. Abd El Munim and S. A. Maged, "Low-Cost Driver Monitoring System Using Deep Learning," in *IEEE Access*, vol. 13, pp. 14151-14164, 2025. https://doi.org/10.1109/ACCESS.2025.3530296

[31] Q. Wu, N. Li, L. Zhang and F. R. Yu, "Driver Drowsiness Detection Based on Joint Human Face and Facial Landmark Localization with Cheap Operations," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 12, pp. 19633 19645, Dec. 2024. https://doi.org/10.1109/TITS.2024.3443832

[32] K. Fujiwara, H. Iwamoto, K. Hori and M. Kano, "Driver Drowsiness Detection Using R-R Interval of Electrocardiogram and Self- Attention Autoencoder," in *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 2956-2965, Jan. 2024. https://doi.org/10.1109/TIV.2023.3308575

[33] D. L. Nguyen, M. D. Putro and K. H. Jo, "Lightweight CNN-Based Driver Eye Status Surveillance for Smart Vehicles," in *IEEE Transactions on Industrial Informatics*, vol. 20, no. 3, pp. 3154-3162, March 2024. https://doi.org/10.1109/TII.2023.3296921

[34] T. Debbarma, T. Pal, A. Saha and N. Debbarma, "HCNNet: A hybrid convolutional neural network for abnormal human driver behaviour detection," *Sādhanā*, vol. 50, no. 1, p. 9, 2025. https://doi.org/10.1007/s12046-024-02656-z

[35] S. Dey, A. Dutta, J. I. Toledo, S. K. Ghosh, J. Lladós and U. Pal, "Signet: Convolutional siamese network for writer independent offline signature verification," *arXiv preprint arXiv:1707.02131*, 2017. https://doi.org/10.48550/arXiv.1707.02131

[36] L. Alzubaidi *et al*., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, pp. 1–74, 2021. https://doi.org/10.1186/s40537-021-00444-8

[37] Y. Albadawi, M. Takruri and M. Awad, "A review of recent developments in driver drowsiness detection systems," *Sensors*, 22, no. 5, 2022. https://doi.org/10.3390/s22052069

[38] Q. Wang, J. Li, X. Tong, and P. M. Atkinson, "TSI Siamnet: A Siamese network for cloud and shadow detection based on time-series cloudy images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 213, pp. 107-123, 2024. https://doi.org/10.1016/j.isprsjprs.2024.05.022

[39] A. Fedele, G. Riccardo and P. Dino, "Explaining Siamese networks in few-shot learning," *Machine Learning*, vol. 113, pp. 7723-7760, 2024. https://doi.org/10.1007/s10994-024-06529-8

[40] S. A. El-Nabi, W. El-Shafai, E. S. M. El-Rabaie, K. F. Ramadan, F. A. Abd El-Samie and S. Mohsen, "Machine learning and deep learning techniques for driver fatigue and drowsiness detection: A review," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 9441-9477, 2024. https://doi.org/10.1007/s11042-023-15054-0

[41] Y. Peng *et al*., "A Multi-Source Fusion Approach for Driver Fatigue Detection Using Physiological Signals and Facial Image" in *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 11, pp. 16614-16624, Nov. 2024. https://doi.org/10.1109/TITS.2024.3420409

[42] D. Gusak, G. Mezentsev, I. Oseledets, and E. Frolov, "RECE: Reduced Cross-Entropy Loss for Large Catalogue Sequential Recommenders," arXiv preprint, *arXiv:2408.02354*, Aug 2024 https://doi.org/10.1145/3627673.3679986

[43] B. Kulambayev, G. Astaubayeva, G. Tleuberdiyeva, J. Alimkulova, G. Nussupbekova and O. Kisseleva, "Deep CNN Approach with Visual Features for Real Time Pavement Crack Detection," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 15, no. 3, 2024. https://doi.org/10.14569/IJACSA.2024.0150333

[44] M. Venkateswarlu and V. Rami Reddy Ch, "DrowsyDetectNet: Driver Drowsiness Detection Using Lightweight CNN With Limited Training Data," in *IEEE Access*, vol. 12, pp. 110476-110491, 2024. https://doi.org/10.1109/ACCESS.2024.3440585

[45] E. Agliari, F. Alemanno, M. Aquaro and A. Fachechi, "Regularization, early-stopping and dreaming: a Hopfield-like setup to address generalization and overfitting," *Neural Networks*, vol. 177, p. 106389, 2024. https://doi.org/10.1016/j.neunet.2024.106389

[46] T. Miseta, A. Fodor and A. Vathy-Fogarassy, "Surpassing early stopping: A novel correlation-based stopping criterion for neural networks," *Neurocomputing*, vol. 567, p. 127028, 2024. https://doi.org/10.1016/j.neucom.2023.127028

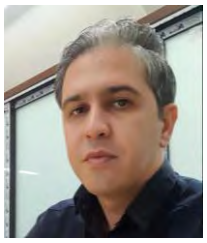
[47] A. Cutkosky, A. Defazio and H. Mehta, "Mechanic: A learning rate tuner," *arXiv preprint, arXiv: 2306.00144*, 2023. https://doi.org/10.48550/arXiv.2306.00144

[48] M. V. Ferro, Y. D. Mosquera, F. J. R. Pena and V. M. D. Bilbao, "Early stopping by correlating online indicators in neural networks," *Neural Networks*, vol. *159*, pp.109-124, 2023. https://doi.org/10.1016/j.neunet.2022.11.035

[49] A. Jami, M. Razzaghpour, H. Alnuweiri and Y. P. Fallah, "Augmented driver behavior models for high fidelity simulation study of crash detection algorithms," *IET Intelligent Transport Systems*, vol. 18, no. 3, pp. 436-449, 2024. https://doi.org/10.1049/itr2.12373

[50] R. Sun *et al.*, "YawnNet: A Visual-Centric Approach for Yawning Detection," In *Proceedings of the International Conference on Multimedia Retrieval*, pp. 1140-1144, 2024. https://doi.org/10.1145/3652583.3657618

[51] M. Simon et al., "EEG alpha spindle measures as indicators of driver fatigue under real traffic conditions," *Clin. Neurophysiol.*, vol. 122, no. 6, pp. 1168–1178, Jun. 2011. https://doi.org/10.1016/j.clinph.2010.10.044



Mahdi Seyfipoor received his B.Sc. degree in Computer Engineering from Khaje Nasir University of technology and M.Sc. in Digital Electronic from Amirkabir University of technology and is currently a Ph.D. candidate in Computer Engineering at University of Tehran. He is Lecturer at the University of Tehran since 2019. His research interests include Computer Vision, Programmable Devices and Real Time Systems.

Mohadeseh Parvizi received her B.S. degree from Hamedan University of Technology and is currently an M.S. student in Artificial Intelligence at Alzahra University.

Siamak Mohammadi (Senior Member, IEEE) received his BSc, MSc and PhD degrees from the University of Paris Sud Orsay, all in electrical engineering. From 1997 to 1999 he was a Research Associate with the Department of Computer Science, University of Manchester, England. In 1999 he moved to Canada and worked in Semiconductor industry in Toronto until 2005. Currently he is an Associate Professor in School of Electrical and Computer engineering, at the University of Tehran, Iran.