Domain-Aware Spatial-Temporal Selective Gated LSTM for Real-Time Driver Drowsiness Detection

Novriadi Antonius Siagian [1], Poltak Sihombing [2,*], Amalia [2], and Ade Candra [2]

[1] Computer Science Doctoral Program, Universitas Sumatera Utara, Medan 20155, Indonesia
[2] Department of Computer Science, Universitas Sumatera Utara, Medan 20155, Indonesia
Email: novriadiantonius@students.usu.ac.id (N.A.S.); poltak@usu.ac.id (P.S.); amalia@usu.ac.id (A.M.); ade_candra@usu.ac.id (A.C.)
*Corresponding author

***Abstract*—Driver drowsiness is a major contributor to road accidents, creating an urgent need for real-time and non-intrusive monitoring systems that operate reliably under practical driving conditions. However, existing vision-based deep learning approaches often suffer performance degradation under illumination variation, facial occlusion, and transient visual noise, limiting their ability to capture subtle temporal manifestations of fatigue during gradual microsleep progression. This study proposes a robust temporal modeling framework for real-time driver drowsiness detection on Internet-of-Things (IoT) edge platforms based on a Domain-Aware Spatial–Temporal Selective-Gated Long Short-Term Memory (SG-LSTM) architecture. In the proposed framework, deep spatial features extracted by ResNet50 are fused with physiologically validated indicators, namely the Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), and head pitch angle, forming a unified 2051-dimensional temporal sequence representation. Two complementary gating mechanisms are introduced to suppress redundant spatial features and emphasize informative time steps during temporal reasoning. Experimental evaluation on the National Tsing Hua University Drowsy Driver Detection (NTHU-DDD) dataset, supported by real driving recordings, demonstrates that the proposed method achieves an accuracy of 90.09% and an F1-score of 0.90, outperforming Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) and vanilla Long Short-Term Memory (LSTM) baselines while maintaining robustness under real-world visual degradation. Deployment on an Orange Pi 3 edge device further confirms the feasibility of real-time inference on resource-constrained hardware. These results indicate that integrating domain-aware physiological cues with selective temporal gating provides an effective and practical solution for real-time driver drowsiness detection in intelligent transportation systems and IoT-edge environments.**

***Keywords*—domain-aware physiological fusion, selective-gated LSTM, spatio-temporal micro-expression analysis, real-time embedded driver monitoring, vision-based drowsiness detection**

Manuscript received November 17, 2025; revised December 17, 2025; accepted January 9, 2026; published April 28, 2026.

I. INTRODUCTION

Amid the increasing complexity of traffic dynamics and the alarming escalation of road accidents, the development of intelligent and non-intrusive safety systems has become a critical research priority. In contemporary transportation infrastructures, the Internet of Things (IoT) enables continuous sensing, real-time monitoring, and automated intervention through integrated tracking and control mechanisms [1, 2]. Within this ecosystem, understanding human behavioral and psychological characteristics is essential, since driver fatigue, distractive behavior, and delayed reflexes remain persistent contributors to collision events [3].

Traditional drowsiness detection approaches rely on manual observation or physiological sensors such as Electroencephalogram (EEG) and Electrooculogram (EOG), which provide high diagnostic precision but are intrusive, difficult to maintain, and impractical for everyday driving environments [4]. Vision-based systems present a non-invasive alternative by analyzing facial expressions and eye movements through standard RGB cameras [5]. However, their performance heavily depends on visual clarity and environmental stability. Real-world conditions introduce illumination variation, eyeglass reflections, shadows, partial occlusions, and motion blur, all of which degrade facial features and reduce recognition reliability [6]. Because drowsiness manifests gradually through subtle multimodal cues, such as prolonged eyelid closure, repeated yawning, and changes in head inclination, extracting discriminative micro-expressions from noisy visual streams remains particularly challenging [7, 8].

Recent advances in deep learning have improved this domain considerably. Long Short-Term Memory (LSTM) networks are capable of modeling long-range temporal dependencies in sequential data and are therefore widely utilized for human activity and driver state analysis [9].

Hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) architectures further enhance representational capacity by combining deep spatial embeddings with temporal learning [10]. Nevertheless, empirical studies show that these models degrade significantly in uncontrolled environments, where spatial textures become unreliable and temporal noise disrupts behavioral transitions [11]. Although physiological cues such as the Eye Aspect Ratio, Mouth Aspect Ratio, and head pitch demonstrate strong correlation with fatigued states [12], most existing approaches incorporate these signals merely as auxiliary features and do not embed them directly within the temporal reasoning mechanism.

Prior studies have explored selective mechanisms and attention-based filtering to suppress irrelevant visual information in sequential deep learning models [13]. However, despite these advancements, existing CNN-LSTM and attention-based approaches predominantly rely on uniform temporal accumulation, where all frames and feature channels are treated equally regardless of their relevance to fatigue progression. This design reduces temporal discriminability and causes brief but critical fatigue-related micro-events, such as short eye closures, transient yawning episodes, and subtle head-nodding movements, to be diluted by redundant spatial features and visually corrupted frames. Despite advancements in drowsiness detection, three major limitations remain evident in existing literature: (1) CNN-LSTM architectures frequently overfit when trained on limited benchmark datasets [14]; (2) simple domain-aware concatenation fails to capture complex spatial–temporal fatigue dynamics [15]; and (3) many state-of-the-art models demand high computational resources, making them unsuitable for IoT-edge deployment with constrained processing and memory capacity [16]. Moreover, recent studies show that prolonged driving sequences introduce landmark drift and temporal discontinuities under illumination variation and motion blur, which further destabilize feature trajectories in low-quality video input [17]. In addition, error accumulation across successive frames can propagate misclassifications when temporal filtering is insufficiently robust, particularly during subtle fatigue progression [18]. Furthermore, most prior studies rely heavily on controlled datasets such as National Tsing Hua University Drowsy Driver Detection (NTHU-DDD) [19], with limited validation on real-world driving environments that exhibit domain shift, illumination variation, and unpredictable occlusion [12].

Taken together, these limitations clearly define the core problem addressed in this work: existing CNN-LSTM-based drowsiness detection frameworks lack selective temporal reasoning, insufficiently integrate domain-aware physiological cues into temporal modeling, and exhibit limited robustness and feasibility under real-world and IoT-edge constraints.

The motivation of this study arises from the gap between the increasing demand for real-time, non-intrusive driver drowsiness monitoring systems and the practical limitations of existing CNN-LSTM and attention-based approaches when deployed under real-world and IoT-edge constraints. While prior studies demonstrate promising accuracy, most rely on uniform temporal accumulation or complex architectural designs, which reduce robustness under visual degradation and limit applicability on resource-constrained embedded devices. Importantly, this work is not motivated by low-level hardware optimization such as frame-rate acceleration or latency benchmarking, but rather by the need for an algorithmic-level solution that enhances temporal discriminability and robustness without increasing architectural complexity.

Motivated by these limitations, this study proposes a domain-aware spatial-temporal Selective-Gated LSTM (SG-LSTM) framework specifically designed for real-time, non-intrusive driver drowsiness detection on embedded IoT hardware. The proposed approach extracts spatial representations using a convolutional neural network and integrates them with physiologically validated indicators, Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), and head pitch, directly within the SG-LSTM gating computation rather than through post-hoc feature concatenation. Two complementary mechanisms are introduced: a channel-selective gate that suppresses redundant or unstable spatial features and a temporal-selective gate that emphasizes frames containing meaningful micro-expressions while filtering noisy temporal segments. The model is evaluated on both the NTHU-DDD benchmark [19] and natural driving recordings [12], enabling assessment under domain shift and real-world noise. Furthermore, the system is optimized for lightweight inference, addressing hardware limitations commonly encountered in IoT-edge deployment [14–16].

The primary contributions of this study are as follows:

1) A novel hybrid CNN-SG-LSTM architecture, in which spatial embeddings from a Convolutional Neural Network are integrated with domain-aware physiological indicators directly at gate-level temporal computation.
2) A dual selective mechanism, channel-selective and temporal-selective gates, that suppresses redundant spatial information and emphasizes discriminative temporal transitions under illumination variation, partial occlusion, motion instability, and low-quality video conditions.
3) An edge-optimized implementation capable of running in real time on resource-limited IoT hardware.
4) A fully operational embedded deployment, validating the feasibility of non-intrusive, real-time drowsiness detection for intelligent transportation and future smart-vehicle safety systems.
5) Extensive evaluation on both NTHU-DDD and external real-driving videos, demonstrating improved robustness, reduced dataset bias, and stronger generalization in uncontrolled environments with illumination changes and natural driver movement.

This study addresses several critical challenges that remain unresolved in prior literature:

(a) Noise-resilient temporal inference. Real-world driving conditions frequently produce unstable visual frames due to illumination shifts, eyeglass glare, occlusion, and camera motion. The proposed SG-LSTM architecture embeds selective gating to preserve informative micro-expressions while suppressing noisy inputs.

(b) Overfitting on limited benchmark datasets. Existing CNN-LSTM approaches often generalize poorly when trained exclusively on controlled datasets. Domain-aware gating and additional evaluation on external real-driving videos are employed to reduce dataset bias and improve robustness.

(c) Modeling complex spatial-temporal interactions. Simple concatenation of features is insufficient to capture the physiological-temporal coupling associated with fatigue progression. SG-LSTM integrates physiological cues into gate computation, enabling deeper representation learning.

(d) IoT-edge deployment under strict resource constraints. Many state-of-the-art models require high computational capacity, making them impractical for in-vehicle embedded systems. This study introduces a lightweight, real-time implementation suitable for low-power edge devices.

The remainder of this article is organized as follows. Section I provides the research background and motivation. Section II reviews related studies. Section III presents the proposed methodology. Section IV reports implementation and evaluation. Section V concludes the paper with key findings and future directions.

II. Literature Review

Research on driver drowsiness detection has advanced rapidly over the past five years, particularly with a paradigm shift from invasive physiological signal-based methods to noninvasive visual approaches. Early solutions relied on EEG and EOG due to their diagnostic precision; however, such methods are impractical for routine driving conditions and unsuitable for real-world deployment [20]. Consequently, a growing research direction focuses on RGB-camera-based visual monitoring combined with facial landmark detection as a more user-friendly alternative. Within this paradigm, Convolutional Neural Networks (CNNs) have been widely adopted for extracting spatial features from the eyes, mouth, and facial regions, while Long Short-Term Memory (LSTM) networks model temporal behaviors such as slow blinking, yawning, or head nodding [12, 21, 22]. Although CNN-LSTM combinations achieve promising results on controlled datasets, their performance often degrades under real-world conditions where illumination variation, eyeglass reflections, and motion blur frequently occur.

To increase robustness, recent studies emphasize architectures capable of selectively retaining salient features while discarding noisy or unstable information [23, 24]. In parallel, domain-aware physiological indicators have gained attention: the Eye Aspect Ratio for blinking and microsleep detection, the Mouth Aspect Ratio for yawning analysis, and head-pitch estimation for nodding behavior [15, 20, 25]. Multi-feature fusion involving Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), and head-pose estimation has shown superior robustness compared to purely CNN-based models, particularly in the presence of visual noise or hardware limitations. Motivated by real-time requirements, several works have transitioned from desktop-class inference to embedded and edge-computing platforms [9, 26, 27], demonstrating the necessity of lightweight yet discriminative models for practical on-vehicle deployment.

Despite these advances, model reliability remains challenged when visual data are degraded by low resolution, occlusion, or nighttime illumination. Studies combining CNN and LSTM on edge platforms confirm that performance deteriorates when video quality declines, underscoring the need for architectures that can stabilize temporal reasoning under noisy conditions [13, 25, 28]. To overcome these limitations, recent research has explored alternative hybrid designs that extend beyond conventional CNN-LSTM pipelines. Hassan *et al.* [29] proposed a Transformer-based model for real-time drowsiness detection, achieving promising results but requiring substantial computational resources—making such solutions impractical for IoT-edge devices.

Efforts to improve accuracy while maintaining efficiency have led to multiple hybrid CNN architectures. Al-Gburi *et al.* [30] introduced EffRes-DrowsyNet, combining EfficientNetB0 with ResNet50 to enhance feature quality; however, model complexity remains a barrier for embedded deployment. Ahmed *et al.* [25] applied VGG16 with data augmentation to reduce overfitting, but VGG-based inference is computationally demanding for low-power environments. To mitigate this, several studies incorporate regularization, early stopping, and lightweight CNN blocks (Conv2D, MaxPooling, Flatten, Dropout, Dense) to balance speed and accuracy.

Advanced attention-based architectures have also emerged. Xiao *et al.* [31] developed ADNet, a deep residual model integrating channel- and spatial-attention modules with augmentation-based training. Although effective, misclassification persists for visually similar behaviors, and the approach depends heavily on labeled multi-label real-driving datasets. Lv *et al.* [32] introduced TSNet using EfficientNet-V2 and a Temporal-Spatial Adaptive Module to extract temporal-spatial features, trained via scheduled stochastic gradient descent. Kumar *et al.* [33] proposed a hybrid model combining Modified InceptionV3 with LSTM, equipped with global average pooling and dropout to improve spatial robustness while modeling temporal dynamics such as eye-closure duration and yawning frequency.

Beyond drowsiness, Farhan *et al.* [34] investigated distraction detection using diverse CNN backbones (ResNet, VGG16, AlexNet, GoogleNet, DenseNet, Xception, MobileNet, and LeNet), with feature optimization via a Self-Adaptive Grass Fibrous Root Optimization (SA-GFRO) algorithm and classification using Modified LSTM. Nevertheless, the approach still suffers from overfitting and sensitivity to input image quality. Li *et al.* [35] employed MobileNetV3_small for binary classification of eye and mouth states, leveraging fine-tuned transfer learning and personalized thresholds. However, performance degrades in nighttime driving and with eyeglasses, while face-detection overhead increases latency. Kim *et al.* [36] proposed an Infrared (IR)-based monitoring system using facial landmarks and solvePnP-based head-pose estimation; still, the method lacks integration with vehicle data such as speed or steering angle.

More complex spatio-temporal architectures have also been studied. Adhithyaa *et al.* [37] introduced a 3D Adaptive State Learning Network utilizing image pyramids and 3D-CNNs with multiplicative fusion. While effective, the multi-module training pipeline is computationally heavy. Bekhouche *et al.* [38] employed pretrained CNNs with transfer learning and sliding temporal windows, yet the lack of explicit temporal-signal modeling (blinks, yawns, nods) limits fine-grained fatigue inference in realistic conditions.

These studies collectively indicate that, despite extensive exploration of CNN-LSTM, attention-based, hybrid, and Transformer-driven architectures, domain-aware physiological signals are predominantly treated as auxiliary inputs or post-hoc feature concatenations. While such strategies enrich feature representations, they do not modify the internal temporal dynamics of recurrent networks, resulting in uniform temporal accumulation that limits sensitivity to brief fatigue-related micro-events. Moreover, most attention mechanisms are applied as external weighting modules rather than being embedded directly within recurrent gate computations, and many recent models prioritize accuracy improvements without explicitly addressing robustness under visual degradation or the computational constraints of IoT-edge deployment [15, 16, 27, 30].

Despite recent advances in vision-based driver drowsiness detection, important research gaps remain. Existing CNN-LSTM and attention-based methods commonly rely on uniform temporal processing and shallow integration of domain-aware physiological cues, which limits their sensitivity to brief but critical fatigue-related micro-events. Moreover, most prior studies are primarily evaluated on controlled datasets, raising concerns about robustness and practical deployment under real-world driving conditions and resource-constrained IoT-edge environments.

From a methodological standpoint, the key superiority of the proposed SG-LSTM lies in its selective temporal reasoning mechanism. Unlike conventional CNN-LSTM and attention-based frameworks that aggregate temporal information uniformly, the proposed approach regulates information flow directly at the recurrent gate level using domain-aware physiological cues, enabling more reliable modeling of short-duration fatigue-related micro-events under noisy visual conditions.

In contrast to existing approaches, the present study introduces a domain-aware Selective-Gated LSTM (SG-LSTM) architecture that integrates physiological cues, EAR, MAR, and head pitch directly into the pre-sigmoid gate computation of the recurrent network. This gate-level fusion enables selective channel-wise and temporal-wise filtering during sequence modeling, allowing the network to preserve discriminative micro-expressions under noisy visual conditions while maintaining computational efficiency suitable for real-time IoT-edge devices. Accordingly, the novelty of this research lies in enabling lightweight, robust, real-time, and non-intrusive drowsiness detection through a selective gating mechanism driven by domain-aware cues.

III. Materials and Methods

With domain-aware signals to provide a more robust framework for accurate and reliable driver drowsiness detection under real-world driving conditions. The input video stream is first processed through face detection and cropping, after which spatial features are extracted using a pre-trained ResNet50 model to obtain high-dimensional representations. In parallel, domain-aware signals are computed to capture critical behavioral indicators. These complementary features are then fused into a unified 2051-dimensional feature representation, as illustrated in Fig. 1. To capture temporal dynamics, a sliding window approach is employed, where each sequence of 30 frames with a stride of 15 is fed into the proposed Selective-Gated LSTM (SG-LSTM). The SG-LSTM introduces two specialized gating mechanisms: the Channel-Selective Gate, which prioritizes relevant spatial features, and the Temporal-Selective Gate, which emphasizes significant temporal dependencies. The final output is classified into two categories: Drowsy and Normal, deployed on an OrangePi IoT device equipped with a buzzer/speaker to deliver real-time alerts, thereby enhancing driving safety.

A. Dataset Description

This study utilizes the National Tsing Hua University Drowsy Driver Detection (NTHU-DDD) dataset [19], developed by the Computer Systems Laboratory at National Tsing Hua University (NTHU), Taiwan. The dataset was specifically designed to capture the dynamics of drivers' facial expressions that reflect varying levels of drowsiness during simulated driving tasks under consistently controlled conditions of lighting, camera positioning, and environment.

The dataset involves 36 participants recorded under five different conditions: without glasses, with glasses, nighttime without glasses, nighttime with glasses, and with sunglasses. These variations are intended to replicate realistic driving scenarios, particularly those involving changes in facial attributes and illumination. The dataset is divided into three primary subsets: 360 video clips from 18 subjects for training, 20 clips from 4 subjects for validation, and 70 clips from 14 subjects for testing. To prevent data leakage, participants are assigned to different subsets with no overlap.

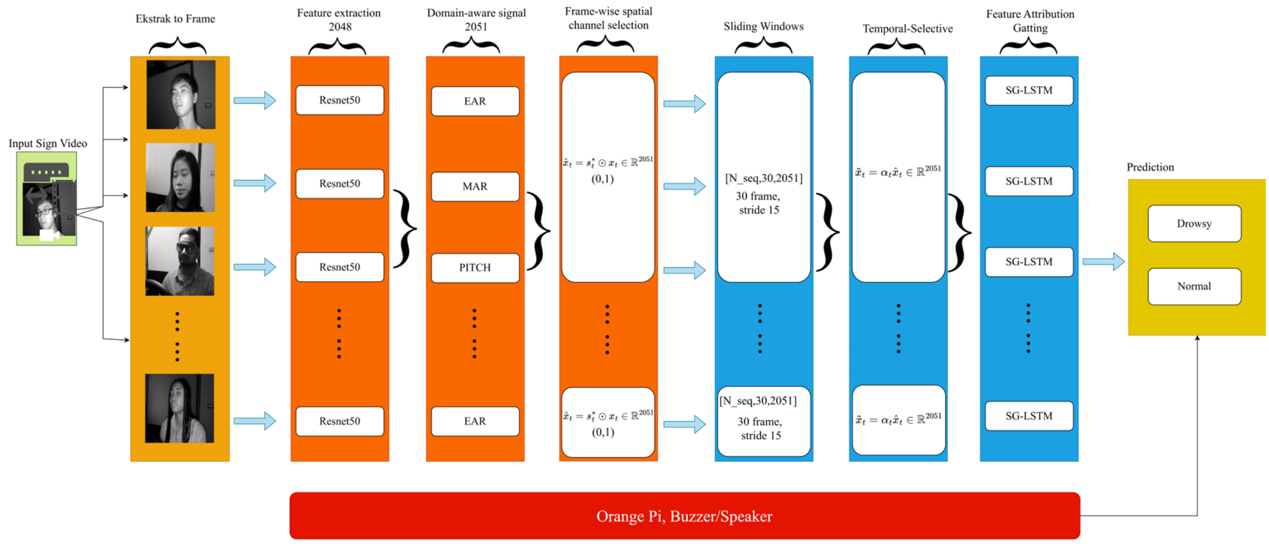

Fig. 1. Proposed drowsiness detection system process.

For the training and validation subsets, frame-level annotations are provided, covering drowsiness status, eye condition, mouth condition, and head posture. In contrast, ground-truth labels for the testing subset are not publicly available. Accordingly, this study employs only the training and validation subsets for model learning and evaluation. Videos recorded under nighttime conditions were captured at 15 fps to accommodate low-light environments, while other recordings were captured at 30 fps. All videos are grayscale, with uniform resolution and no accompanying audio.

Beyond the public dataset, additional real-time primary data were collected using an embedded system equipped with an RGB camera as the main visual sensor, as shown in Fig. 2. This primary dataset enables evaluation of the proposed SG-LSTM framework under real-world operational conditions, providing an assessment of inference consistency and system reliability when deployed on edge devices. Thus, the study not only validates the model using a standardized benchmark dataset but also provides empirical evidence of its effectiveness in real-world scenarios.

Fig. 2. National Tsing Hua University (NTHU-DDD) video data collection in 5 conditions.

B. *Data Preprocessing*

The preprocessing stage begins with frame extraction from the input videos, where each recording at 30 Frames Per Second (FPS) is converted into a sequence of images per second. This process transforms raw video data into a more structured unit of analysis, making it suitable for subsequent processing steps. Once individual frames are obtained, face detection is performed using MediaPipe Face Detection. The detected facial regions are then cropped to obtain the relevant Region of Interest (ROI), while non-facial areas are discarded to reduce visual noise. The cropped faces are subsequently resized to a standardized resolution of 224×224 pixels, and pixel intensity values are normalized to the range [0, 1], ensuring stability and consistency during processing by the Convolutional Neural Network (CNN).

C. *Domain-Aware Feature Extraction*

In detecting driver drowsiness, visual representations derived from CNNs often fall short in capturing subtle and transient physiological indicators [39]. To address this limitation, the present study integrates domain-aware signals that are physiologically validated to correlate with drivers' levels of alertness. These domain-aware features are selected for their robustness against variations in illumination, eyeglass reflections, and spatial noise, while simultaneously representing genuine biological phenomena that are difficult to capture through conventional visual approaches alone [27, 35, 40]. Three key indicators are employed: the Eye Aspect Ratio (EAR), the Mouth Aspect Ratio (MAR), and the Pitch Angle, each providing critical information associated with microsleep, yawning, and head-nodding behaviors, respectively.

1) *Eye Aspect Ratio (EAR)*

The Eye Aspect Ratio is a metric used to quantify the degree of eye openness based on the Euclidean distances between upper and lower eyelid landmarks. EAR has been widely adopted for detecting microsleep or partial eye closure, as individuals experiencing drowsiness typically exhibit increased blink duration and frequency, often followed by prolonged eyelid closure. Mathematically, EAR is defined as [41]:

$$\mathrm{EAR} = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2 \cdot \|p_1 - p_4\|} \tag{1}$$

This equation defines the Eye Aspect Ratio (EAR), which is used to quantify eye openness and detect blink/microsleep events. The points $p_1, \ldots, p_6$ denote the eye landmarks obtained from facial landmark detection, where, $\|p_2 - p_6\|$ *and* $\|p_3 - p_5\|$ represent the vertical eyelid distances and, $\|p_1 - p_4\|$ represents the horizontal eye width. The denominator $2 \cdot \|p_1 - p_4\|$ normalizes EAR to reduce inter-subject variation in eye size, enabling consistent thresholding and temporal modeling.

2) Mouth Aspect Ratio (MAR)

The Mouth Aspect Ratio, on the other hand, measures the degree of mouth opening and serves as a dominant indicator of yawning behavior. Yawning is recognized as an early symptom of drowsiness that is visually observable and has a sufficiently long duration to be captured temporally [15]. The MAR is mathematically expressed as:

$$\mathrm{MAR} = \frac{\|p_{14} - p_{18}\| + \|p_{15} - p_{17}\| + \|p_{13} - p_{19}\|}{2 \cdot \|p_{12} - p_{16}\|} \tag{2}$$

This equation defines the Mouth Aspect Ratio (MAR), which is used to capture mouth opening dynamics associated with yawning. The landmarks $p_{12}, \ldots, p_{19}$ correspond to mouth keypoints, where the numerator aggregates multiple vertical lip distances and $\|p_{12} - p_{16}\|$ epresents the mouth width used for normalization. MAR provides an interpretable physiological cue that complements visual CNN features when texture quality is degraded.

3) Pitch Angle (Head Pose)

Identifying behavioral cues such as head nodding or tilting, which are characteristic indicators of microsleep, necessitates the application of head-pose estimation. The pitch angle is derived using a head-pose estimation method based on the rotation matrix of the facial coordinate system [42]. Specifically, the pitch component is calculated as the rotational displacement of the head relative to the camera along the horizontal axis and is mathematically expressed as follows:

$$\theta_{\text{pitch}} = \arctan\, 2\left(-R_{31}, \sqrt{R_{11}^2 + R_{21}^2}\right) \tag{3}$$

This equation computes the head pitch angle θ_{pitch}, which is used to detect head-nodding behavior indicative of fatigue. The terms R_{ij} are elements of the 3D rotation matrix R estimated from head-pose computation (e.g., PnP-based pose estimation). The arctan $2(\cdot)$ formulation ensures a stable angle estimate across quadrants, while $\sqrt{R_{11}^2 + R_{21}^2}$ acts as a normalization term for projecting rotation onto the horizontal plane.

D. Spatial Feature Extraction

In the spatial feature extraction stage, this study employs ResNet50 as the backbone network (Fig. 4). ResNet50 is a Convolutional Neural Network (CNN) architecture introduced with the concept of residual learning, in which residual blocks enable the construction of very deep networks while mitigating the vanishing gradient problem. With 50 layers and over 23 million parameters, ResNet50 is capable of producing rich and discriminative feature representations from input images [43]. In this work, the model is used in its pre-trained form on the ImageNet dataset to leverage transfer learning, allowing each processed facial frame to be represented as a 2048-dimensional feature vector.

$$Y = F(X, \{W_i\}) + X \tag{4}$$

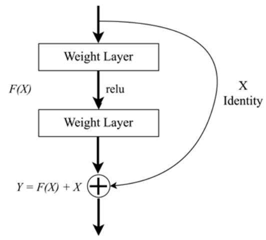

Fig. 3. Residual block in ResNet50.

This residual formulation defines how ResNet learns spatial representations while mitigating vanishing gradients, as illustrated in Fig. 3. Here, X denotes the input feature map to a residual block $F(X, \{W_i\})$ is the residual function (a stack of convolution, normalization, and nonlinearity) parameterized by weights W_i, and Y is the block output. The skip connection "$+X$" preserves low-level information and stabilizes training, producing a robust 2048-dimensional embedding after global pooling for each frame.

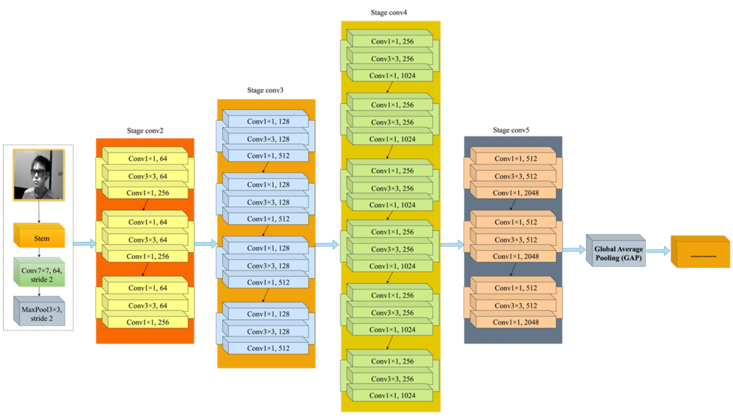

Fig. 4. ResNet-50 architecture as a pre-trained feature extractor.

E. Feature Fusion

The feature fusion stage aims to combine spatial information extracted by the CNN with domain-aware signals, thereby creating a more holistic representation of the driver's state. In this study, each frame processed by ResNet50 produces a 2048-dimensional feature vector. At the same time, three physiologically grounded domain-aware features are derived from facial landmarks, forming an additional 3-dimensional vector.

These two representations are merged through a concatenation operation, resulting in a unified 2051-dimensional feature vector. Before fusion, both feature sets are normalized using either z-score normalization or min-max scaling to place them on comparable ranges. This prevents any single feature type from dominating the learning process and ensures that domain-aware information is preserved rather than being overshadowed by the larger magnitude of CNN-derived features.

$$F_t = [X_t \oplus g_t] \quad (5)$$

This equation defines the fusion of spatial and domain-aware features at time step t. The vector, $X_t \in \mathbb{R}^{2048}$ represents the CNN (ResNet50) spatial embedding extracted from frame t, and $g_t = \left[EAR_t, MAR_t, \theta_{\text{pitch},t}\right] \in \mathbb{R}^3$ denotes the domain-aware physiological indicators. The operator $\oplus$ denotes concatenation, yielding the unified representation $F_t \in \mathbb{R}^{2051}$ used as the input sequence for temporal modeling.

F. Temporal Windowing

Each frame produces a unified vector $F_t \in \mathbb{R}^{2051}$. For temporal modeling, we form a sequence of duration T frames with stride s: $\mathrm{F}_{t:t+T-1} \in \mathbb{R}^{T\times 2051}$. The number of windows from the video along the L frame is calculated with:

$$N = \left\lfloor \frac{L-T}{s} \right\rfloor + 1 \quad (6)$$

This equation computes the number of sliding temporal windows used for sequence construction. Here, L denotes the total number of frames (or feature vectors) in a video, T_s is the window length (sequence length) in frames, and N is the number of resulting sequences when using a stride of 1. This windowing step ensures the temporal model receives fixed-length inputs for training and inference.

G. Selective-Gated LSTM

The proposed architecture integrates the per-frame representation $F_t \in \mathbb{R}^{2051}$ comprising 2048 CNN features obtained after global average pooling $\oplus$ 3 domain-aware features: EAR, MAR, and Pitch) with two selective gating mechanisms, namely, the Channel-Selective Gate and the Temporal-Selective Gate, prior to entering the LSTM cell. The purpose of this design is to suppress non-informative channels while emphasizing temporal segments that are most relevant to drowsiness episodes, thereby enabling a more robust modeling of spatial-temporal dynamics in real-time scenarios on edge devices.

1) Long Short-Term Memory(LSTM)

Artificial neural networks based on Recurrent Neural Networks (RNNs) [44], were developed to address the key limitation of classical RNNs, namely, their inability to retain information over long sequences due to the vanishing gradient problem. Long Short-Term Memory (LSTM), first introduced by Hochreiter and Schmidhuber in 1997, was specifically designed to process sequential data such as text, audio, video, and time-series signals.

The structure of an LSTM comprises three principal components: *input gate* (Γ_u), *forget gate* (Γ_f), *output gate* (Γ_o). At each time step, the input (x_t) is received by the LSTM cell, which regulates long-term memory through the *cell state* (c_t) and produces the *hidden state* (h_t) The activation function tanh is employed to manage sequential information across time, allowing the network to capture both short-term dependencies and long-range temporal dynamics.

$$\text{Forget Gate } f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (7)$$

$$\text{Input Gate } i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (8)$$

$$\text{Output Gate } o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (9)$$

$$\text{Candidate Cell } \tilde{C}_t = tan\,h(W_c x_t + U_c h_{t-1} + b_c) \quad (10)$$

$$\text{Cell State Update } C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (11)$$

$$\text{Hidden State } h_t = o_t \odot \tanh(C_t) \quad (12)$$

Eqs. (7)–(12) describe the standard LSTM update used to model temporal dependencies in the fused sequence. The input x_t denotes the feature vector at time step t. (in this work, $x_t \in \mathbb{R}^{2051}$), and h_{t-1} is the previous hidden state encoding past temporal context. The gates f_t, i_t, and o_t control memory retention, information injection, and output exposure, respectively, using the sigmoid function $\sigma(\cdot)$ to produce values in $[0, 1]$. The candidate cell $\tilde{C}_t$ proposes new content through $\tanh(\cdot)$, while $\odot$ denotes element-wise multiplication. The trainable matrices W and U and biases b are learned during training to optimize drowsiness classification performance.

2) Channel-selective gate

For each frame, the fused vector $F_t \in \mathbb{R}^{2051}$ is weighted channel-wise in order to suppress noisy channels and emphasize informative ones. The initial channel activations are derived through a pre-sigmoid linear projection, followed by a sigmoid function to regulate the gating process:

$$s_t = \sigma(W_c F_t + b_c) \in (0, 1)^{2051} \quad (13)$$

This equation computes the channel-selective gating vector s_t, which is used to suppress redundant or noisy feature channels in the fused representation F_t. The weight matrix W_c and bias b_c are trainable parameters that map F_t to a per-dimension gate, and $\sigma(\cdot)$ constrains the gate values to (0, 1) to enable soft selection.

The scalar τ_c defines the minimum gate activation used to enforce controlled sparsity. It determines how aggressively small gate values are suppressed, thereby reducing sensitivity to unstable channels caused by illumination variation or occlusion.

To achieve controlled selectivity, this study applies soft-thresholding with a predefined threshold. $\tau_c = 0.1$

$$s_t^* = \frac{max(s_t - \tau_c, 0)}{1 - \tau_c} \in [0, 1]^{2051} \quad (14)$$

This equation applies thresholding and rescaling to obtain s_t^*. Values below τ_c are set to zero via max $(\cdot, 0)$, while the division by $1 - \tau_c$ rescales the remaining activations back to $[0, 1]$. This operation increases selectivity by filtering weakly informative channels.

The thresholded mask is then applied through element-wise multiplication with the fused vector to produce the selectively retained features:

$$\hat{x}_t = s_t^* \odot F_t \in \mathbb{R}^{2051} \quad (15)$$

This equation produces the refined input $\hat{x}_t$. by applying the channel gate s_t^* to the fused vector F_t The resulting x_t is then forwarded to the temporal stage, ensuring that the LSTM receives a denoised sequence representation.

3) Temporal-selective gate

In the Temporal-Selective Gating (TSG) stage, each frame within the temporal window is assigned a scalar weight that depends on past context, thereby preserving causality. The TSG generates a per-time weight $\alpha_t \in (0, 1)$ which amplifies informative frames (e.g., microsleep episodes or head-nodding) while attenuating non-salient ones. To ensure efficiency on resource-constrained edge devices, the weights are implemented as scalars and then broadcast across the 2051 channels:

Pre-activation:

$$a_{\alpha,t} = w_\alpha^\top h_{t-1} + b_\alpha \quad (16)$$

This equation computes a scalar relevance score $a_{\alpha,t}$ used to assess the importance of the current time step. The vector w_α and bias b_α are learnable parameters, while h_{t-1} provides temporal context from prior frames.

gate temporal:

$$\alpha_t = \sigma(a_{\alpha,t}) \in (0, 1) \quad (17)$$

This equation maps the relevance score to a temporal gate α_t in $(0, 1)$, enabling soft selection of informative time steps and down-weighting of corrupted or redundant frames.

feature weighting:

$$\tilde{x}_t = \alpha_t \hat{x}_t \in \mathbb{R}^{2051} \quad (18)$$

This equation applies the temporal gate α_t to the feature vector $\hat{x}_t$, amplifying time steps associated with salient micro-events (e.g., short eye closures) while attenuating noisy segments, thereby improving temporal discriminability.

H. Selective Gate Long Short-Term Memory (SG-LSTM)

The proposed Selective-Gated Long Short-Term Memory (SG-LSTM) architecture extends the classical LSTM by introducing a pre-sigmoid gating mechanism within the input gate pathway. This modification is designed to provide more fine-grained control over the spatial-temporal information entering the memory cell. Unlike conventional LSTMs, which directly accumulate the input signal x_t and the previous hidden state h_{t-1} through a linear transformation followed by a sigmoid activation, the SG-LSTM incorporates an additional early selection layer that enables information filtering at an earlier stage.

The core innovation lies in the pre-sigmoid gating mechanism. In contrast to standard LSTMs that execute the input gate equation directly, the SG-LSTM introduces a pre-sigmoid layer that functions as an adaptive filter. This stage is further enhanced by integrating an intermediate transformation through the selective signal g_t and an auxiliary module, ensuring that only the most relevant spatial-temporal features are emphasized before entering the memory cell. The mechanism is formally expressed as follows:

$$\tilde{x}_t = \alpha_t \hat{x}_t \quad (19)$$

where α_t serves as the temporal attention coefficient that regulates the scale of the input prior to the main sigmoid activation. Subsequently, the input gate is computed as:

$$i_t^{SG} = \sigma(W_i x_t + U_i h_{t-1} + V_i(\sigma(W_\alpha x_t + U_\alpha h_{t-1} + b_\alpha) \odot s_t) + b_i) \quad (20)$$

This equation, $W_i x_t$ represents the spatial mapping of the input, $U_i h_{t-1}$ integrates temporal memory, while V_i conveys the additional signal derived from the pre-sigmoid gating combined with the domain-aware signal s_t. The parameter b_i continues to serve as a bias term, stabilizing the activation. Through this design, the SG-LSTM is able to perform adaptive selection of relevant signals before they enter the memory cell.

This formulation can be further expressed in a more compact form as:

$$i_t^{SG} = \sigma(W_i \tilde{x}_t + U_i h_{t-1} + V_i g_t + b_i) \quad (21)$$

In this equation, $W_i \tilde{x}_t$ represents the extraction of spatial information that has already been filtered, while $U_i h_{t-1}$ integrates the temporal dynamics from the previous hidden state. The term $V_i g_t$ serves as an adaptive modulating signal derived from the pre-sigmoid gating, enriching the feature selection process. Meanwhile, b_i functions as a bias term to stabilize the activation.

SG-LSTM enables fine-grained gating over the input dimensions, allowing the model to suppress irrelevant visual disturbances or domain signals before the final sigmoid activation regulates the proportion of information admitted into the memory cell. This mechanism ensures that the driver drowsiness detection system focuses attention on critical signals such as eye dynamics, mouth movements, and head orientation, while minimizing the influence of irrelevant features caused by lighting noise or individual attribute variations.

To further enhance clarity and reproducibility, the complete algorithmic procedure of the proposed Spatial-Temporal Selective-Gated LSTM is summarized in the following pseudo-code.

Algorithm: Pseudo-Code of the Spatial-Temporal Selective Gated LSTM

Input: a sequence $X = [x_1, \ldots, x_T]$, initial states h_0, c_0
Output: logits $\hat{y}$, probabilities $\hat{p}$, and hidden states {ht}
begin
Set $h_0 \leftarrow 0$, $c_0 \leftarrow 0$.
for $t = 1$ to T do
  Compute s_t as Eq. (13).
  Compute $\hat{x}_t$ as Eq. (15).
  Compute $a_{\alpha,t}$ as Eq. (16).
  Compute α_t as Eq. (17).
  Compute $\tilde{x}_t$ as Eq. (18).
  Compute i_t^{SG} as Eqs. (19)–(20).
  Update f_t, o_t, $\tilde{C}_t$ (standard LSTM).
  Update C_t as.
  Update h_t.
end for
Compute $\hat{y}$.
Compute $\hat{p}$.
return $\hat{y}$, $\hat{p}$.
end

I. Hyperparameter Model

During training, the model was optimized with a batch size of 16, balancing gradient stability with memory efficiency for both CPU and edge-device environments. The training was conducted for 50 epochs to ensure convergence while mitigating the risk of overfitting. A conservative learning rate of 1×10^{-4} was adopted to provide stable weight updates, and the Adam optimizer was employed, leveraging both momentum and adaptive learning. This configuration is particularly well-suited to the complexity of sequential modelling tasks.

J. Model Evaluation

Model evaluation in this study was conducted using standardized performance metrics, namely accuracy, precision, recall, and F1-score. All of these metrics are derived from the confusion matrix [3], which illustrates the distribution of model predictions relative to the actual class labels.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{22}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{23}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{24}$$

$$F1-\text{Score} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{25}$$

Accuracy measures the overall proportion of correct predictions (the sum of true positives and true negatives) against the total number of test samples, thereby providing a general overview of model performance. However, this metric may become biased in the presence of class imbalance. Precision assesses the correctness of positive predictions, i.e., the proportion of samples predicted as drowsy that are indeed drowsy, making it sensitive to false positives. Recall evaluates the coverage of positive case detection, indicating how many actual drowsy cases were successfully identified by the model, and is thus sensitive to false negatives. The F1-score, defined as the harmonic mean of precision and recall, summarizes the trade-off between the two and is particularly valuable when class distributions are imbalanced or when both false positives and false negatives carry critical consequences.

Together, these four metrics provide a concise yet comprehensive assessment of the model's discriminative capability and its susceptibility to the most relevant types of misclassification in driver safety scenarios. By combining them, the evaluation not only emphasizes global accuracy but also highlights the system's consistency in avoiding critical prediction errors. This ensures that the proposed model achieves not only strong statistical performance but also practical reliability and relevance for real-world deployment.

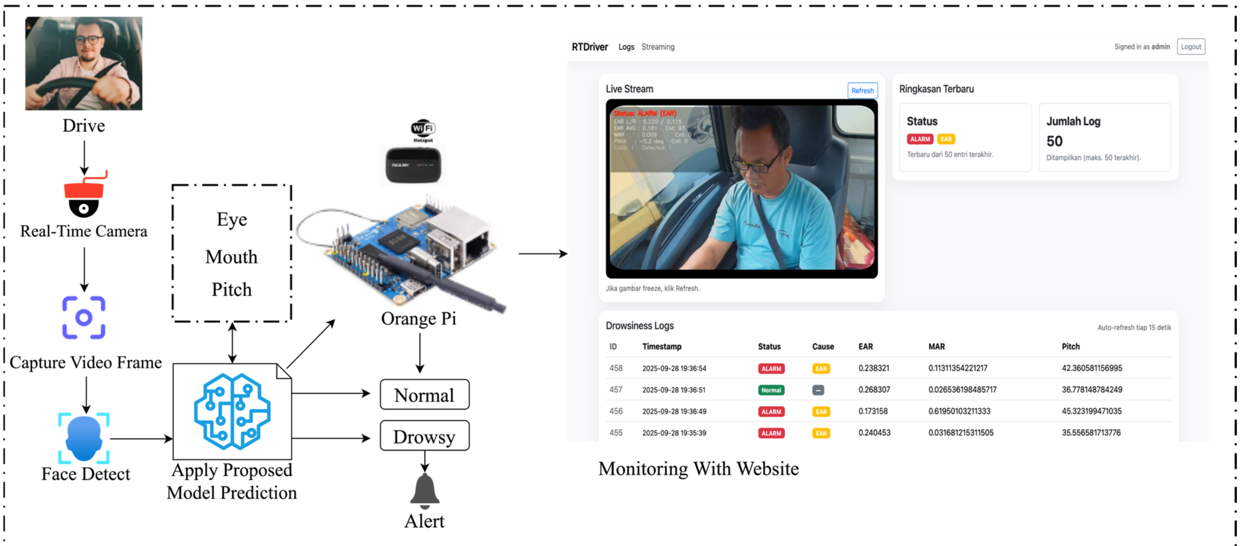

Fig. 5. Proposed real-time monitoring architecture.

K. Deployment on IoT Devices to Embedded System

The embedded system in this study was implemented on the Orange Pi 3 (4 GB RAM), an open-source edge-computing board selected for its energy efficiency, wireless integration, and hardware versatility, as shown in Fig. 5. Powered by a quad-core Cortex-A53 processor (1.5 GHz) and a Mali G31 GPU, the device provides adequate parallel computing for spatial feature extraction and SG-LSTM inference. With 4 GB LPDDR4 memory and expandable microSD storage, it ensures stable sequential data handling during operation. Connectivity features include Wi-Fi 5, Bluetooth 5.0, Gigabit Ethernet, HDMI, USB, UART, and GPIO, supporting seamless integration with sensors and actuators. In this configuration, an RGB camera captured driver facial data, a buzzer served as an auditory alert, and a push button enabled manual control. With compact 50×55 mm dimensions and low power consumption, the Orange Pi 3 delivers a practical embedded platform for real-time, non-invasive driver monitoring, demonstrating the feasibility of SG-LSTM deployment on edge devices and its strong potential to improve transportation safety in real-world settings

L. Camera Calibration

Camera calibration is a critical stage in video-based visual monitoring systems, particularly in the context of real-time driver state detection. The primary objective of this process is to ensure that the camera remains consistently aligned with the driver's face, thereby producing accurate geometric projections. In this study, calibration was performed to minimize image distortion and to guarantee the precision of domain-aware signal estimation, including the Eye Aspect Ratio, Mouth Aspect Ratio, and head pitch angle, all of which are highly dependent on camera position and orientation. The system retains only those camera configurations that capture the driver's full face within the frame and achieve a high landmark detection accuracy (>95%).

IV. RESULT AND DISCUSSION

This section presents the experimental results and performance analysis of the proposed model. The discussion covers the dataset, preprocessing, training outcomes, evaluation metrics, comparative analysis with other methods, and deployment on edge devices. The focus is placed on the effectiveness of the Domain-Aware Spatial-Temporal SG-LSTM architecture in detecting driver drowsiness in real time and its relevance to IoT-based safety systems.

B. Data Distribution Analysis

The initial analysis focused on assessing the distribution of frames across the two target classes. As illustrated in Fig. 6, the dataset consists of 176,847 frames labeled as Normal and 171,976 frames labeled as Drowsy, indicating a relatively balanced distribution. Such balance is critical to mitigate class bias and to ensure the robustness of the learning process.

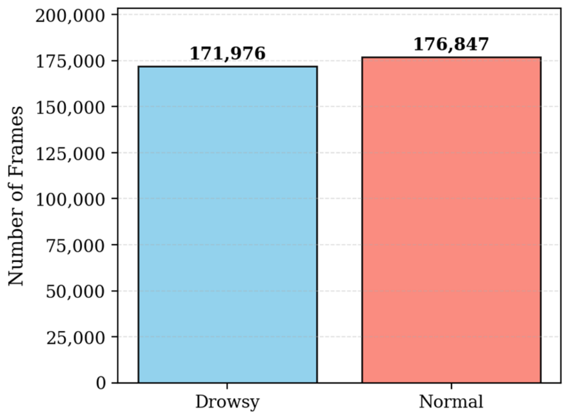

Fig. 6. Distribution of frames across conditions.

C. Dataset after Preprocessing and Domain-Aware Feature Extraction

After confirming a balanced distribution, the preprocessing stage was continued with domain-aware feature extraction, which included the Eye Aspect Ratio, Mouth Aspect Ratio, and head pitch. This process produced a structured dataset that was ready for model training. Table I presents a representative subset of the preprocessed data, linking frame indices, class labels, and the corresponding feature values. This representation provides a detailed illustration of the data format employed in the experiments.

TABLE I. DATASET AFTER PREPROCESSING AND DOMAIN-AWARE FEATURE EXTRACTION

frame_name	frame_index	timestamp	label	EAR	MAR	pitch
frame_0000_0p00.jpg	0	0	0	0.4699882531	0.03096355918	414.884.464
frame_0001_0p10.jpg	1	00.01	0	0.4722165677	0.02264757751	4.115.383.534
frame_0002_0p20.jpg	2	00.02	0	0.4676296324	0.02015585793	4.133.348.483
frame_0003_0p30.jpg	3	00.03	0	0.475501622	0.0213510784	4.231.295.742
frame_0004_0p40.jpg	4	00.04	0	0.4724252648	0.01765102188	4.156.630.017
frame_0005_0p50.jpg	5	00.05	0	0.4649537176	0.03122143363	4.191.844.369
…..	…..	…..	…..	…..	…..	…..
…..	…..	…..	…..	…..	…..	…..
frame_0720_72p00.jpg	720	72	1	0.4291940339	0.01442290004	4.384.484.002

D. Model Training Results

The training results of the Domain-Aware Spatial-Temporal SG-LSTM model are presented in Figs. 7 and 8, which respectively illustrate the loss and accuracy curves over 50 epochs. As shown in Fig. 7, the training loss consistently decreased, while the validation loss stabilized after approximately the 20th epoch. This trend indicates that the model successfully achieved convergence without exhibiting significant overfitting.

Fig. 7. Lost over epochs.

In Fig. 8, the accuracy curves demonstrate a gradual improvement for both training and validation data, eventually stabilizing above 90% after the 40th epoch. This performance highlights the effectiveness of integrating domain-aware signals (EAR, MAR, and pitch) along with the selective gating mechanism in enhancing the model's generalization ability. Consequently, the proposed architecture is not only efficient during the learning process but also capable of maintaining consistent performance on unseen data.

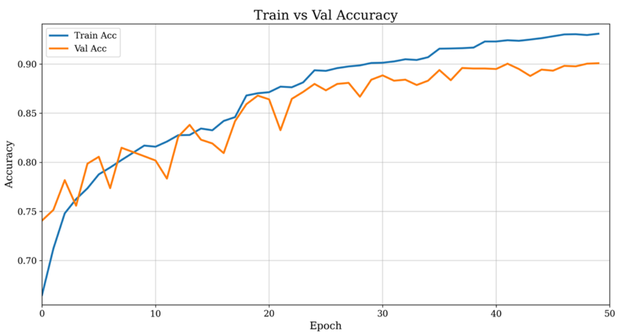

Fig. 8. Accuracy over epochs.

E. Evaluation Metrics

The evaluation of the proposed Domain-Aware Spatial-Temporal SG-LSTM model was conducted using both visual and quantitative approaches to ensure a comprehensive understanding of its performance. The confusion matrix (Fig. 9) provides a direct visualization of classification outcomes, showing that the majority of samples were accurately classified on the main diagonal. Specifically, 813 normal cases and 851 drowsy cases were correctly identified, with only minor errors occurring in the form of 119 normal samples misclassified as drowsy and 64 drowsy samples incorrectly predicted as normal. This distribution reflects the model's stability and strong ability to discriminate between the two classes with relatively low error rates, which is critical in safety-critical applications such as driver monitoring.

Complementing this, the classification report (Table II) offers a detailed breakdown of performance metrics. For the Normal class, the model achieved a precision of 0.9270, a recall of 0.8723, and an F1-score of 0.8988, indicating its reliability in minimizing false positives while maintaining sensitivity. For the Drowsy class, the recall of 0.9029 underscores the model's capacity to detect the majority of true drowsiness cases, which is particularly important for real-world applications where failure to detect fatigue could have severe consequences. The overall accuracy of 90.09%, supported by consistent macro and weighted averages around 0.90, confirms the balanced trade-off between precision and recall, further validating the robustness of the model.

These findings highlight the effectiveness of incorporating domain-aware features, Eye Aspect Ratio, Mouth Aspect Ratio, and head pitch, combined with the selective gating mechanism, which collectively enhance temporal sequence modeling and generalization. The strong alignment between visual evidence from the confusion matrix and numerical indicators from the classification report reinforces the conclusion that the proposed architecture is both accurate and reliable. Beyond demonstrating technical robustness, these results emphasize the model practical relevance for real-time driver drowsiness detection in intelligent transportation systems and IoT-based safety solutions.

F. Comparative Analysis

To validate the improvements introduced by the proposed architecture, the performance of the Domain-Aware Spatial-Temporal SG-LSTM was compared against a baseline Vanilla LSTM model using the same dataset and experimental configuration. As summarized in Table III, the SG-LSTM consistently outperformed the classical LSTM across all key evaluation metrics.

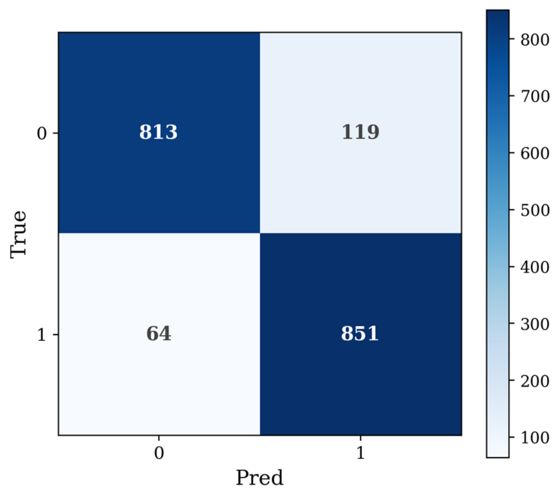

Fig. 9. Confusion matrix.

TABLE II. PERFORMA ACCURACY, PRECISION, RECALL, F1-SCORE

Label	Precision	Recall	F1-Score	Support
Normal	0.9270	0.8723	0.8988	932
Drowsy	0.8773	0.9301	0.9029	915
Accuracy			0.9009	1847
Macro Avg	0.9022	0.9012	0.9009	1847
Weighted Avg	0.9024	0.9009	0.9009	1847

TABLE III. COMPARISON OF ACCURACY, PRECISION, RECALL, AND F1 SCORE

Model	Precision	Recall	F1-Score	Support
Vanilla LSTM	0.8779	0.8771	0.8771	1847
SG-LSTM	0.9024	0.9012	0.9009	1847

The conventional LSTM demonstrates reasonable capability in modeling temporal dependencies; however, it treats all input features and time steps uniformly, which may limit its ability to preserve subtle fatigue-related micro-events under visual noise. In contrast, the proposed SG-LSTM incorporates domain-aware features together with selective gating mechanisms that regulate information flow at both channel and temporal levels.

This selective gating allows the model to suppress redundant or unstable visual information while emphasizing fatigue-relevant cues such as eye closure dynamics, yawning behavior, and head-nodding patterns. As a result, the SG-LSTM achieves more stable and discriminative temporal representations than the conventional LSTM, leading to improved robustness and reliability in real-time driver drowsiness detection scenarios.

G. Performance Comparison with Baseline and Previous Studies

Despite the advantages demonstrated by the proposed approach, the comparative analysis in Table IV also reveals several inherent limitations that merit consideration. First, although the selective-gated architecture improves robustness under visual degradation, the framework still relies on reliable facial region detection and landmark estimation for domain-aware signal extraction. In extreme scenarios involving severe occlusion, unconventional camera angles, or complete landmark failure, the quality of EAR, MAR, and head-pitch estimation may deteriorate, potentially affecting temporal inference stability.

Second, while the proposed model reduces dependence on deep spatial backbones compared to CNN-heavy architectures, its performance remains influenced by the quality of the initial spatial representations extracted by ResNet50. Although this backbone offers a favorable balance between accuracy and efficiency, further reduction of computational complexity may be required for deployment on ultra-low-power embedded platforms with stricter memory and energy constraints.

Third, as reflected by the reliance of all compared methods in Table IV on the NTHU-DDD dataset, the current evaluation is bounded by a semi-controlled benchmark environment. While real-time demonstrations partially mitigate this limitation, broader validation on large-scale naturalistic driving datasets with higher behavioral diversity, sensor variability, and environmental complexity is necessary to fully assess long-term generalization.

TABLE IV. COMPARISON WITH BASELINES

Model	Accuracy	Feature	Dataset	Key Strength (Efficacy)	Edge / Real-time Suitab
Dlib + HOPEnet + SVM/RF/XGBoost [4]	83.96%	Facial Features, Eyes, Mouth, Head Pose	NTHU	Interpretable features and low training complexity	Moderate (lightweight ML but unstable landmarks)
RF-DCM+LSTMs [5]	89.42%	Face, Eyes, Mouth, Glabella	NTHU	Strong spatial feature recalibration and improved accuracy	Limited (heavy CNN blocks)
2DCNN+TSAM [6]	89.42%	Face, eye	NTHU	Attention enhances salient spatial–temporal regions	Moderate (not optimized for low-power edge)
Tiny-YOLO+ Resnet50+SVM [7]	86.74%	Face Detection	NTHU	Robust face detection under moderate motion	Limited (face detector + deep CNN)
Proposed Model	90.09%	Eyes, Mouth, pitch angle	NTHU	Selective temporal gating preserves micro-events; robust under visual degradation; lightweight temporal modeling	designed for real-time IoT-edge deployment

Finally, the proposed framework focuses exclusively on vision-based and physiologically derived cues. Although selective temporal gating improves robustness relative to existing vision-only approaches, integrating complementary modalities such as vehicle dynamics, inertial measurements, or wearable physiological sensors may further enhance resilience under conditions where visual information becomes unreliable.

H. Computational Complexity and Runtime Analysis

The proposed SG-LSTM extends a conventional LSTM by introducing channel-selective and temporal-selective gating operations composed of linear projections and element-wise activations. These additional operations do not modify the asymptotic time complexity of the recurrent module and introduce only a marginal increase in per-epoch computation compared to a standard LSTM.

During inference, the selective gating mechanism operates on fixed-length feature vectors and does not require additional convolutional or attention-based processing. As a result, the inference cost remains comparable to that of a conventional CNN-LSTM pipeline while avoiding the computational overhead associated with deeper CNN backbones or transformer-based temporal models. This design enables stable real-time inference on resource-constrained IoT-edge devices.

I. Real-Time Demonstration of the Proposed Model

To validate its practical applicability, the proposed model was implemented on an embedded system and evaluated under real-time conditions. Figs. 9 and 10 presents a representative snapshot of the system output, showcasing the model's capability to directly identify signs of driver drowsiness from live video input. This demonstration confirms the feasibility of the developed architecture for deployment on resource-constrained embedded platforms.

Real-time demonstrations, as presented in Figs. 9 and 10, further examined the effectiveness of the proposed approach. Using the NTHU-DDD dataset (Fig. 9), the system successfully identified multiple indicators of drowsiness, including eye closure events, yawning patterns, and head pose deviations (pitch). These outputs validate the capability of the proposed architecture to generalize across controlled experimental conditions while maintaining robust classification of both drowsy and normal states.

In addition, a test drive scenario (Fig. 11) was conducted to evaluate the system in a naturalistic driving environment. The results demonstrate that the model consistently recognized fatigue-related behaviors, such as prolonged yawning and nodding, while effectively distinguishing them from non-fatigued conditions. This real-world demonstration confirms the practical relevance of the proposed method and highlights its readiness for deployment on embedded platforms. Collectively, these findings reinforce the robustness, adaptability, and real-time applicability of the system in ensuring driver safety.

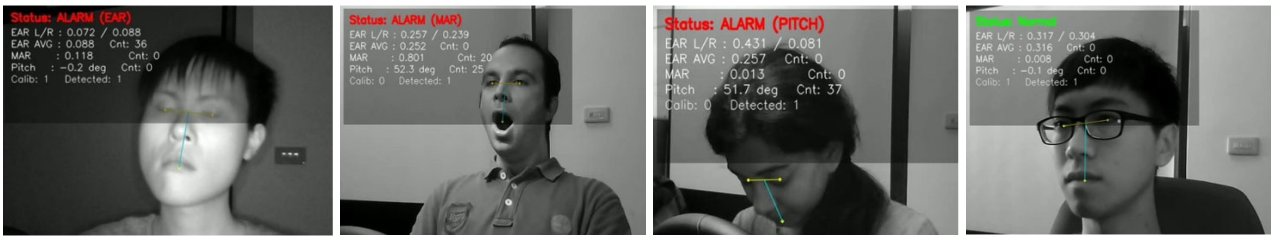

Fig. 10. Demonstration of the proposed dataset NTHU-DDD.

Fig. 11. Demonstration of the proposed test drive.

V. Conclusion

This study presents a domain-aware spatial-temporal selective-gated LSTM framework for real-time driver drowsiness detection, combining deep spatial representations with physiologically grounded cues within a selective temporal modeling architecture. By incorporating eye aspect ratio, mouth aspect ratio, and head pitch into channel- and temporal-level gating mechanisms, the proposed approach enables effective modeling of subtle fatigue-related behaviors while maintaining computational efficiency.

Experimental results on the NTHU-DDD dataset, together with real-time deployment on an embedded IoT platform, demonstrate that the proposed framework delivers reliable detection performance and is suitable for operation in resource-constrained environments. These outcomes highlight the practical relevance of selective temporal reasoning for intelligent transportation systems, where robustness and real-time responsiveness are critical for improving road safety.

The primary contribution of this work lies in introducing a selective-gated recurrent architecture that integrates domain-aware physiological information directly into temporal modeling without increasing asymptotic computational complexity. From a theoretical standpoint, this framework demonstrates that gate-level selective temporal modulation provides a more reliable representation of fatigue-related micro-events than conventional CNN–LSTM formulations. From a practical standpoint, the proposed SG-LSTM is well suited for real-time, non-intrusive driver monitoring on embedded IoT-edge platforms, where robustness to visual degradation and stable temporal inference are essential.

Future research will focus on extending the framework to multimodal sensing configurations, validating its performance on large-scale naturalistic driving datasets, and systematically comparing the proposed SG-LSTM with other advanced LSTM variants, as well as further optimizing hardware-software co-design for ultra-low-power edge devices. Through these efforts, the proposed approach has strong potential to advance driver monitoring technologies and support safer and more intelligent transportation systems.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Author Contributions

N.A.S. conceived the research, designed the methodology, and conducted the experiments. P.S., as the main supervisor, I provided overall supervision, conceptual guidance, and significant contributions to the refinement of the manuscript. A.M. and A.C., as co-supervisors, offered substantial feedback, supported the validation of results, and contributed actively to the research framework and system implementation. All authors contributed to writing, reviewed the content, and approved the final version of the manuscript.

Acknowledgment

The authors would like to express their profound gratitude to the supervisor and co-supervisor for their invaluable guidance, constructive feedback, and continuous support throughout the course of this doctoral research. The authors also extend their sincere appreciation to Universitas Sumatera Utara for providing academic support and a conducive research environment that enabled the successful completion of this study. Furthermore, the authors are deeply grateful to the test drivers at PT. Raja Sukses Transindo for their participation and cooperation during the real-world test drive, which provided essential validation for the proposed system.

References

[1] S. Chen, Z. Wang, and W. Chen, "Driver drowsiness estimation based on factorized bilinear feature fusion and a long-short-term recurrent convolutional network," *Inf.*, vol. 12, no. 1, pp. 1–15, Jan. 2021. doi: 10.3390/info12010003

[2] X. Zhan, W. Wu, L. Shen, W. Liao, Z. Zhao, and J. Xia, "Industrial internet of things and unsupervised deep learning enabled real-time occupational safety monitoring in cold storage warehouse," *Saf. Sci.*, vol. 152, Aug. 2022. doi: 10.1016/j.ssci.2022.105766

[3] F. Ali, A. Ali, M. Imran, R. A. Naqvi, M. H. Siddiqi, and K. S. Kwak, "Traffic accident detection and condition analysis based on social networking data," *Accid. Anal. Prev.*, vol. 151, Mar. 2021. doi: 10.1016/j.aap.2021.105973

[4] S. C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *Nature Research*, Dec. 2020. doi: 10.1038/s41746-020-00341-z

[5] M. H. Ahmed *et al.*, "Real-time driver depression monitoring for accident prevention in smart vehicles," *IEEE Access*, vol. 12, pp. 79838–79850, 2024. doi: 10.1109/ACCESS.2024.3407361

[6] J. E. Hoffmann, H. G. Tosso, M. M. D. Santos, J. F. Justo, A. W. Malik, and A. U. Rahman, "Real-time adaptive object detection and tracking for autonomous vehicles," *IEEE Trans. Intell. Veh.*, vol. 6, no. 3, pp. 450–459, Sep. 2021. doi: 10.1109/TIV.2020.3037928

[7] A. Natarajan, V. Krishnasamy, and M. Singh, "Device-free human motion detection using single link WiFi channel measurements for building energy management," *IEEE Embed. Syst. Lett.*, vol. 15, no. 3, pp. 153–156, Sep. 2023. doi: 10.1109/LES.2022.3214333

[8] J. N. Cheltha, C. Sharma, D. Prashar, A. A. Khan, and S. Kadry, "Enhanced human motion detection with hybrid RDA-WOA-based RNN and multiple hypothesis tracking for occlusion handling," *Image Vis. Comput.*, vol. 150, Oct. 2024. doi: 10.1016/j.imavis.2024.105234

[9] Y. Yi, H. Zhang, W. Zhang, Y. Yuan, and C. Li, "Fatigue working detection based on facial multifeature fusion," *IEEE Sens. J.*, vol. 23, no. 6, pp. 5956–5961, Mar. 2023. doi: 10.1109/JSEN.2023.3239029

[10] Y. P. Huang, S. Kshetrimayum, and C. T. Chiang, "Object-based hybrid deep learning technique for recognition of sequential actions," *IEEE Access*, vol. 11, pp. 67385–67399, 2023. doi: 10.1109/ACCESS.2023.3291395

[11] Y. Wu, H. Sheng, Y. Zhang, S. Wang, Z. Xiong, and W. Ke, "Hybrid motion model for multiple object tracking in mobile devices," *IEEE Internet Things J.*, vol. 10, no. 6, pp. 4735–4748, Mar. 2023. doi: 10.1109/JIOT.2022.3219627

[12] A. C. Phan, T. N. Trieu, and T. C. Phan, "Driver drowsiness detection and smart alerting using deep learning and IoT," *Internet of Things (Netherlands)*, vol. 22, Jul. 2023. doi: 10.1016/j.iot.2023.100705

[13] R. Huang, Y. Wang, Z. Li, Z. Lei, and Y. Xu, "RF-DCM: Multi-granularity deep convolutional model based on feature recalibration and fusion for driver fatigue detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 630–640, Jan. 2022. doi: 10.1109/TITS.2020.3017513

[14] W. Guo and L. Jing, "Toward low-cost passive motion tracking with one pair of commodity Wi-Fi devices," *IEEE J. Indoor Seamless Position. Navig.*, vol. 1, pp. 39–52, Jun. 2023. doi: 10.1109/jispin.2023.3287508

[15] X. X. Tang and P. Y. Guo, "Fatigue driving detection methods based on drivers wearing sunglasses," *IEEE Access*, vol. 12, pp. 70946–70962, 2024. doi: 10.1109/ACCESS.2024.3394218

[16] J. Tang *et al.*, "Attention-guided multiscale convolutional neural network for driving fatigue detection," *IEEE Sens. J.*, vol. 24, no. 14, pp. 23280–23290, 2024. doi: 10.1109/JSEN.2024.3406047

[17] H. Huang, L. Zhao, and Y. Wu, "An IoT and machine learning enhanced framework for real-time digital human modeling and motion simulation," *Comput. Commun.*, vol. 212, pp. 78–89, Dec. 2023. doi: 10.1016/j.comcom.2023.09.024

[18] E. Perkins, C. Sitaula, M. Burke, and F. Marzbanrad, "Challenges of driver drowsiness prediction: The remaining steps to implementation," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1319–1338, Feb. 2023. doi: 10.1109/TIV.2022.3224690

[19] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016. doi: 10.1109/JIOT.2016.2579198

[20] M. Adil Khan, T. Nawaz, U. S. Khan, A. Hamza, and N. Rashid, "IoT-based non-intrusive automated driver drowsiness monitoring framework for logistics and public transport applications to

enhance road safety," *IEEE Access*, vol. 11, pp. 14385–14397, 2023. doi: 10.1109/ACCESS.2023.3244008

[21] T. Xu *et al.*, "E-key: An EEG-based biometric authentication and driving fatigue detection system," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 864–877, Apr. 2023. doi: 10.1109/TAFFC.2021.3133443

[22] L. Mou *et al.*, "Isotropic self-supervised learning for driver drowsiness detection with attention-based multimodal fusion," *IEEE Trans. Multimed.*, vol. 25, pp. 529–542, 2023. doi: 10.1109/TMM.2021.3128738

[23] L. Mou *et al.*, "Driver stress detection via multimodal fusion using attention-based CNN-LSTM," *Expert Syst. Appl.*, vol. 173, Jul. 2021. doi: 10.1016/j.eswa.2021.114693

[24] P. Liu, H. L. Chi, X. Li, and J. Guo, "Effects of dataset characteristics on the performance of fatigue detection for crane operators using hybrid deep neural networks," *Autom. Constr.*, vol. 132, Dec. 2021. doi: 10.1016/j.autcon.2021.103901

[25] M. I. B. Ahmed *et al.*, "A deep-learning approach to driver drowsiness detection," *Safety*, vol. 9, no. 3, Sep. 2023. doi: 10.3390/safety9030065

[26] P. Thakur, S. Goel, and E. Puthooran, "Edge AI enabled IoT framework for secure smart home infrastructure," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 3369–3378. doi: 10.1016/j.procs.2024.04.317

[27] V. Owen and N. Surantha, "Computer vision-based drowsiness detection using handcrafted feature extraction for edge computing devices," *Appl. Sci.*, vol. 15, no. 2, Jan. 2025. doi: 10.3390/app15020638

[28] D. Salem and M. Waleed, "Drowsiness detection in real-time via convolutional neural networks and transfer learning," *J. Eng. Appl. Sci.*, vol. 71, no. 1, Dec. 2024. doi: 10.1186/s44147-024-00457-z

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

[30] 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, p. 3711, Jun. 2025. doi: 10.3390/s25123711

[31] W. Xiao, H. Liu, Z. Ma, and W. Chen, "Attention-based deep neural network for driver behavior recognition," *Futur. Gener. Comput. Syst.*, vol. 132, pp. 152–161, Jul. 2022. doi: 10.1016/j.future.2022.02.007

[32] X. Lv, G. Zheng, H. Zhai, K. Zhou, and W. Zhang, "Driver fatigue detection method based on temporal—spatial adaptive networks and adaptive temporal fusion module," *Comput. Electr. Eng.*, vol. 119, Oct. 2024. doi: 10.1016/j.compeleceng.2024.109540

[33] V. Kumar, S. Sharma, and Ranjeet, "Driver drowsiness detection using modified deep learning architecture," *Evol. Intell.*, vol. 16, no. 6, pp. 1907–1916, Dec. 2023. doi: 10.1007/s12065-022-00743-w

[34] H. M. Farhan, A. K. Türkben, and R. A. S. Naseri, "Optimal feature tuning model by variants of convolutional neural network with LSTM for driver distract detection in IoT platform," *Knowl. Inf. Syst.*, vol. 67, no. 6, pp. 5151–5186, Jun. 2025. doi: 10.1007/s10115-025-02342-4

[35] X. Li, H. Lin, J. Du, and Y. Yang, "Computer vision-based driver fatigue detection framework with personalization threshold and multi-feature fusion," *Signal, Image Video Process.*, vol. 18, no. 1, pp. 505–514, Feb. 2024. doi: 10.1007/s11760-023-02733-6

[36] D. Kim, H. Park, T. Kim, W. Kim, and J. Paik, "Real-time driver monitoring system with facial landmark-based eye closure detection and head pose recognition," *Sci. Rep.*, vol. 13, no. 1, Dec. 2023. doi: 10.1038/s41598-023-44955-1

[37] N. Adhithyaa, A. Tamilarasi, D. Sivabalaselvamani, and L. Rahunathan, "Face positioned driver drowsiness detection using multistage adaptive 3D convolutional neural network," *Inf. Technol. Control*, vol. 52, no. 3, pp. 713–730, 2023. doi: 10.5755/j01.itc.52.3.33719

[38] S. E. Bekhouche, Y. Ruichek, and F. Dornaika, "Driver drowsiness detection in video sequences using hybrid selection of deep features," *Knowledge-Based Syst.*, vol. 252, Sep. 2022. doi: 10.1016/j.knosys.2022.109436

[39] Y. Jebraeily, Y. Sharafi, and M. Teshnehlab, "Driver drowsiness detection based on convolutional neural network architecture optimization using genetic algorithm," *IEEE Access*, vol. 12, pp. 45709–45726, 2024. doi: 10.1109/ACCESS.2024.3381999

[40] S. M, P. Sathwik, K. B. P. Reddy, S. S. M, and S. P. Raja, "Machine learning algorithm-based driver drowsiness detection system," *Int. J. Intell. Transp. Syst. Res.*, 2025. doi: 10.1007/s13177-025-00530-8

[41] J. Cai, X. Liao, J. Bai, Z. Luo, L. Li, and J. Bai, "Face fatigue feature detection based on improved D-S model in complex scenes," *IEEE Access*, vol. 11, pp. 101790–101798, 2023. doi: 10.1109/ACCESS.2023.3314665

[42] Y. Zhang and Q. Tang, "Accelerating autonomy: An integrated perception digital platform for next generation self-driving cars using faster R-CNN and DeepLabV3," *Soft Comput.*, vol. 28, no. 2, pp. 1633–1652, Jan. 2024. doi: 10.1007/s00500-023-09510-0

[43] V. V. K. Reddy, S. Cherukuri, K. Vanaparla, and L. R. Avula, "Deep feature extraction for fashionable fabrics: Using ResNet50, MobileNet, and CNN," in *Lecture Notes in Networks and Systems*, Springer Science and Business Media Deutschland GmbH, 2025, pp. 417–429. doi: 10.1007/978-981-97-7178-3_36

[44] Y. Dai, J. Wei, and F. Qin, "Recurrent Neural Network (RNN) and Long Short-Term Memory neural network (LSTM) based data-driven methods for identifying cohesive zone law parameters of nickel-modified carbon nanotube reinforced sintered nano-silver adhesives," *Mater. Today Commun.*, vol. 39, Jun. 2024. doi: 10.1016/j.mtcomm.2024.108991