Driver Drowsiness Prediction Using CNN-LSTM Model Based on Facial Expression and Eye Movement

Annisa Aprilia Putri Sakri[1]*, Angga Rusdinar[2], Giva Andriana Mutiara[3], Periyadi[4]
[1,2,3,4]Electrical Engineering Study Program, Faculty of Electrical Engineering, Telkom University, Indonesia

Corresponding Author's e-mail: *aasgoas@gmail.com*

ARMADA
JURNAL PENELITIAN MULTIDISIPLIN

e-ISSN: 2964-2981

ARMADA : Jurnal Penelitian Multidisiplin

https://ejournal.45mataram.ac.id/index.php/armada

Vol. 04, No. 08 Agustus, 2026

Page: 3889-3903

DOI:
https://doi.org/10.55681/armada.v4i8.3341

Article History:

Received: Juni 08, 2026

Revised: Juli 08, 2026

Accepted: Agustus 18, 2026

Abstract : Driver fatigue and drowsiness represent primary institutional catalysts for fatal highway traffic anomalies worldwide. This comprehensive investigation introduces an adaptive, multi-task deep learning architecture merging Convolutional Neural Networks and Long Short-Term Memory configurations to dynamically evaluate driver states through localized facial expressions and non-invasive ocular metrics. Utilizing MediaPipe FaceMesh, the framework maps 468 distinct landmark parameters under fluctuating illumination constraints to monitor regional variations across the eyes and mouth. The quantitative metrics are extracted by formulating real-time computations of the Eye Aspect Ratio and Mouth Aspect Ratio. Spatiotemporal feature representation is accomplished using a pre-trained ResNet50V2 feature extractor integrated with a 128-unit recurrent LSTM layer to process sequences across a 20-frame context window. The multi-branch dense layer concurrently outputs predictive status conditions for drowsiness, yawning frequency, and categorical facial expressions. System validation metrics indicate an overall classification accuracy of 88% for structural drowsiness tracking and 90% for yawning anomalies. This system is a "proof of concept" designed to be implemented for drivers to reduce traffic accidents caused by driver fatigue.

Keywords : CNN-LSTM; Driver Drowsiness; MediaPipe FaceMesh; ResNet50V2; Eye Aspect Ratio.

Abstrak: *Kelelahan dan kantuk pengemudi merupakan faktor utama yang memicu kecelakaan lalu lintas fatal di jalan raya di seluruh dunia. Penelitian komprehensif ini memperkenalkan arsitektur pembelajaran mendalam multitugas adaptif yang menggabungkan Convolutional Neural Networks dan konfigurasi Long Short-Term Memory untuk mengevaluasi kondisi pengemudi secara lebih dinamis melalui ekspresi wajah lokal dan metrik okular noninvasif. Dengan menggunakan MediaPipe FaceMesh, kerangka kerja ini memetakan 468 parameter titik wajah berbeda pada kondisi pencahayaan yang berubah-ubah untuk memantau variasi regional pada mata dan mulut. Metrik kuantitatif diperoleh melalui perhitungan waktu nyata Eye Aspect Ratio dan Mouth Aspect Ratio. Representasi fitur spasiotemporal dilakukan menggunakan ekstraktor fitur ResNet50V2 pralatih yang diintegrasikan dengan lapisan LSTM berulang 128 unit untuk memproses urutan dalam jendela konteks 20 frame. Lapisan padat bercabang secara bersamaan menghasilkan prediksi kondisi kantuk, frekuensi menguap, dan kategori ekspresi wajah. Metrik validasi sistem menunjukkan akurasi klasifikasi keseluruhan*

sebesar 88% untuk pelacakan kantuk struktural dan 90% untuk anomali menguap. Sistem ini merupakan "bukti konsep" yang dirancang untuk diterapkan pada pengemudi guna mengurangi kecelakaan lalu lintas akibat kelelahan pengemudi.

Kata kunci: *CNN-LSTM; Kantuk Pengemudi; MediaPipe FaceMesh; ResNet50V2; Eye Aspect Ratio.*

INTRODUCTION

The architectural integrity of traffic safety on high-speed intercity highways represents a foundational pillar of modern transportation infrastructure and intelligent civil engineering, directly impacting the socio-economic stability, public safety, and operational efficiency of contemporary urban societies (Tumminello et al., 2025). With the exponential escalation of automotive volume and transit density across regional shipping and commuting corridors, the statistical probability of fatal road anomalies and high-impact traffic collisions has shown a persistent upward trend globally (Mani & Goniewicz, 2023). Systematic investigations compiled from national traffic accident registries universally indicate that human operational errors constitute the primary institutional catalyst for severe highway disasters (Bhatta et al., 2025). Specifically, these disasters are triggered by a drastic decline in driver vigilance caused by cumulative physical fatigue, extended operational hours, and psychological drowsiness (Fu et al., 2024).

Prolonged driving periods, particularly during nocturnal schedules or monotonous long-haul transits, frequently force the human nervous system into the highly hazardous phenomenon of microsleep (Fu et al., 2024). Microsleep is characterized as an involuntary, transient episode of complete unconsciousness accompanied by a total cessation of motor responsiveness, lasting anywhere from a fraction of a second up to thirty consecutive seconds (Zaky et al., 2023). When a heavy commercial or passenger vehicle is operating at high speeds, even a momentary two-second lapse in driver awareness translates to dozens of meters traveled completely out of control, invariably culminating in catastrophic, fatal head-on or rear-end collisions (Balazadeh Meresht et al., 2023).

To mitigate this critical vulnerability, early preventive engineering frameworks heavily relied on invasive wearable apparatuses such as electroencephalogram (EEG) headbands, electrocardiogram (ECG) chest straps, or smart biometric wristbands Tactical-Grade Wearables and Authentication Biometrics (Agiomavritis & Karanasiou, 2026). However, these frameworks are consistently rejected by commercial fleet operators and daily commuters due to their highly intrusive physical movement restrictions, sensor placement discomfort, and high maintenance overhead, rendering non-intrusive, vision-based monitoring platforms highly imperative for real-world automotive integration (Visconti et al., 2025).

To overcome the inherent physical limitations of intrusive monitoring methods, computer-vision-based non-invasive monitoring frameworks integrated directly into vehicle cabins have emerged as a principal domain in intelligent transportation systems and artificial intelligence research (Khanal et al., 2022). By deploying an internal dashboard camera equipped with near-infrared or standard optical sensors, subtle and discrete physiological indicators can be monitored dynamically in real-time (Selvaraju et al., 2022). These indicators include localized variations in micro-facial expressions, shifts in blink frequency, the duration of eyelid closure, and the mechanical stretching of the jaw or oral cavity during yawning.

However, the practical deployment of conventional computer vision algorithms and shallow machine learning architectures confronts severe operational bottlenecks when transitioning from controlled laboratory settings to the unpredictable environment of a vehicle cabin (Jafari et al., 2024). These legacy systems demonstrate an extreme sensitivity to rapidly fluctuating cabin illumination constraints such as intense sunlight glare, shadows from roadside trees, or complete darkness during night driving. They are also sensitive to variable head-pose rotations, facial occlusions from steering movements, and the extensive anatomical diversity of individual human faces. Such environmental and spatial instabilities frequently induce erroneous landmark tracking

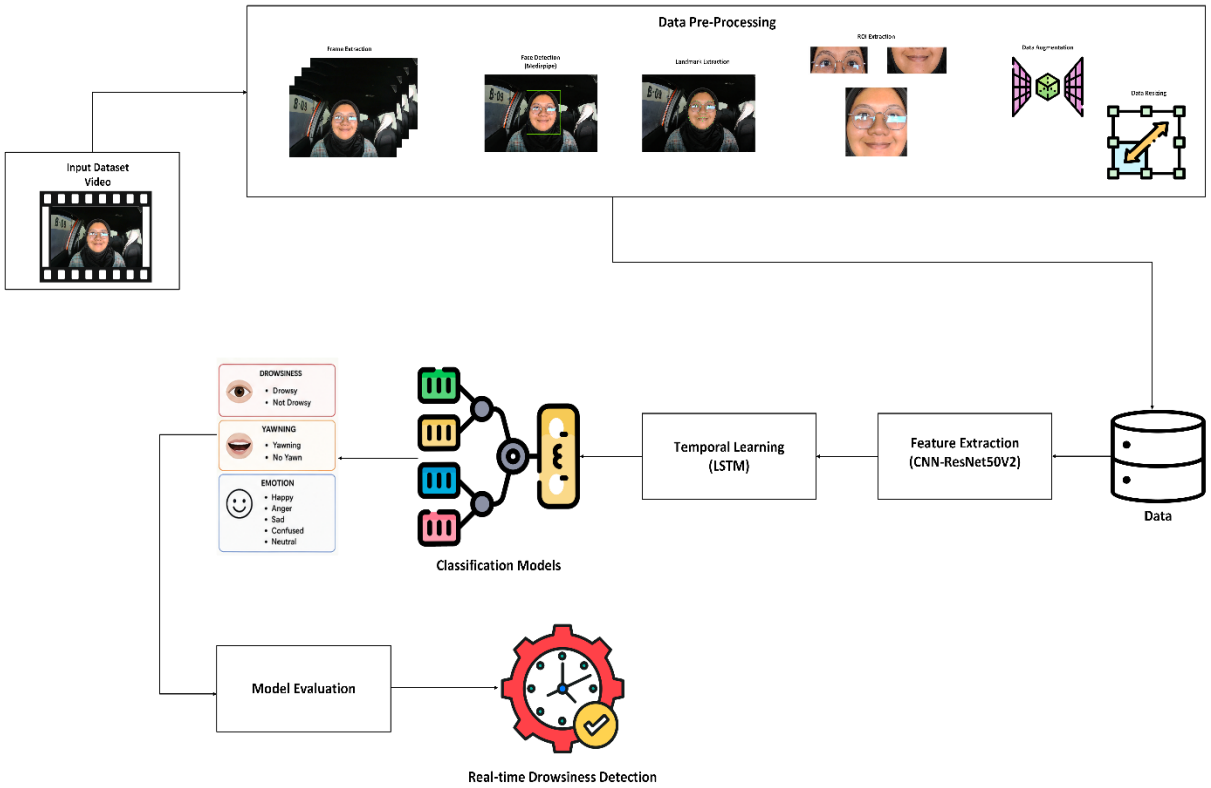
and catastrophic drift in bounding boxes, which significantly degrades the operational reliability of early warning platforms, leading to unacceptable rates of false alarms or missed detections under real-world driving conditions (Asperti & Filippini, 2023).

Recent paradigm shifts in deep learning engineering offer a highly robust alternative through the integration of unified spatiotemporal networks that process both visual features and time-series transitions concurrently. Hybrid neural architectures that blend Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) recurrent blocks have proven highly resilient for processing continuous video sequences captured from dynamic environments (Darwich & Bayoumi, 2024). In these configurations, the konvolusional network component specializes in extracting dense, high-level spatial feature representations and localized patterns from individual video frames (Xiao et al., 2023). Concurrently, the subsequent recurrent LSTM layer tracks the underlying temporal dependencies, sequence transitions, and rate-of-change metrics across successive frames to accurately identify the continuous behavioral shift of a driver from a fully alert state to profound drowsiness (El-Nabi et al., 2024).

This spatiotemporal representation is substantially augmented and accelerated by utilizing the advanced MediaPipe FaceMesh framework (Cao et al., 2024). This library acts as a lightweight, front-end structural penjejak capable of mapping 468 distinct three-dimensional facial coordinate variables with low computational overhead and millimeter-level architectural precision. This structural mapping allows for the precise calculation of localized geometric metrics, such as the Eye Aspect Ratio (EAR) and Mouth Aspect Ratio (MAR), directly from the coordinate points without requiring heavy raw image transformations. While extensive literature covers vision-based fatigue tracking, a notable research gap persists in the domain of multi-task deep learning frameworks that concurrently process multiple heterogeneous behavioral anomalies under unexpected computational constraints (Li et al., 2024). Prior investigations predominantly decouple the tasks of binary drowsiness classification, yawn frequency tracking, and multi-class emotion recognition into isolated, independent networks (Dileep Kumar et al., 2025). This division multiplies the total parameter count, training times, and resource demands on embedded automotive hardware. Furthermore, existing models are typically optimized and validated assuming ideal, uninterrupted training convergence cycles. As a result, there is a distinct lack of empirical evaluations regarding how truncated training cycles or premature resource cutoffs affect the distinct feature headers of a multi-task network (Mao et al., 2024).

This research directly addresses these computational challenges and engineering gaps by developing and comprehensively evaluating an adaptive, multi-task CNN-LSTM deep learning architecture for non-intrusive driver state prediction (AL-Quraishi et al., 2024). The primary objective of this study is to construct a unified network capable of concurrently predicting binary drowsiness (*drowsy_output*), identifying yawning cycles (*yawn_output*), and classifying five distinct categorical facial expressions (*emotion_output*) using a sequence length window of 20 video frames. The core contribution of this work lies in the architectural optimization of a pre-trained ResNet50V2 spatial feature extractor coupled with a 128-unit recurrent LSTM block under rigid data-balancing constraints executed via the RandomUnderSampler algorithm. This study provides a unique empirical analysis of multi-head network resilience by evaluating the model's performance when the training phase is prematurely terminated by a cloud infrastructure interruption (*KeyboardInterrupt*) on Google Colab. The quantitative findings show that, although the temporal weight adjustment stopped abruptly due to GPU limitations, the spatial transfer learning features retain an exceptional classification capacity for distinct expressions this provides important conceptual design insights for the implementation of the Driver Safety System.

RESEARCH METHOD



The system begins with the process of downloading a dataset consisting of the classes Drowsiness, Yawning, and Emotion. Since the dataset consists of videos, the system performs frame extraction to convert the videos into a collection of image frames. Next, each frame is processed using MediaPipe FaceMesh to detect faces and generate facial landmarks. Based on these landmarks, the system extracts Regions of Interest (ROI) that include the face, eyes, and mouth, then calculates the Eye Aspect Ratio (EAR) and Mouth Aspect Ratio (MAR) as key features to determine the condition of the eyes and mouth. The extraction results are then used to label the data and are stored as metadata.

Before the training process, the data was organized into sequences consisting of 20 consecutive frames. The data was then preprocessed by resizing it to 112 × 112 pixels, rescaling the pixel values to the range of 0–1, and applying data augmentation to increase the variety of the training data. This step was designed to ensure that the model had consistent data and was more robust to various lighting conditions and face positions. During the modeling phase, each frame is processed using ResNet50V2 to extract spatial features; then, features from the entire sequence are analyzed by a Long Short-Term Memory (LSTM) network to capture changes in facial expressions over time. The LSTM output is fed into a multi-task classification model that simultaneously generates three predictions: Drowsy/Not Drowsy, Yawning/No Yawn, and Emotion. These predictions are used to determine the driver's level of alertness and can serve as the basis for issuing early warnings if signs of drowsiness are detected.

Spatial Landmark Tracking and Mathematical Formulations

The front-end pipeline utilizes the MediaPipe FaceMesh framework to establish a coordinate system over the driver's facev. This tool maps a total of 468 distinct three-dimensional facial landmarks, which are normalized across varying head positions and cabin angles (Tsai et al., 2026). Within this structural mesh, specific spatial regions corresponding to the left eye, right eye, and oral cavity are tracked. The coordinates for the left eye are determined by tracking landmark indices 362, 385, 387, 263, 373, and 380, whereas the right eye is mapped using indices 33, 160, 158, 133, 153, and 144. To track eyelid changes over successive intervals, the Eye Aspect Ratio metric

is calculated for each eye. The calculation of the Eye Aspect Ratio metric is mathematically formulated in Equation 1.

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

The variables in Equation 1 represent specific cartesian coordinate pairs of the eyelid boundaries, where p_1 and p_4 denote the horizontal corner landmarks of the eye structure, while p_2, p_3, p_5, and p_6 represent the vertical coordinates mapping the upper and lower eyelids. An average value is derived from both eyes to determine the final threshold. A drop below 0.21 identifies an involuntary closure event caused by drowsiness. Simultaneously, the oral cavity is tracked by grouping landmark indices 13, 14, 78, and 308 to compute the Mouth Aspect Ratio metric, which evaluates yawning anomalies. The computation of the Mouth Aspect Ratio metric is defined in Equation 2.

$$MAR = \frac{|p_{top} - p_{bottom}|}{|p_{left} - p_{right}|}$$

The variables in Equation 2 represent the vertical and horizontal spatial extensions of the driver's lips, where the numerator calculates the vertical distance between the upper landmark p_{top} and lower landmark p_{bottom}, while the denominator calculates the horizontal width between the left corner p_{left} and right corner p_{right} of the mouth. A threshold value exceeding 0.19 identifies an open mouth state associated with yawning. Based on these coordinates, a bounding box isolates the eyes with a strict 3:1 aspect ratio, reshaping the final image array to 112 × 112 pixels with three channels to serve as input for the deep neural network.

Hybrid CNN-LSTM Architecture and Multi-Head Design

The core neural architecture is constructed as a deep spatiotemporal network, stacking a convolutional pipeline with a recurrent sequence analyzer (John et al., 2023). The spatial feature extraction head uses the ResNet50V2 model, which leverages transfer learning from the ImageNet database (Montañez & Alon, 2025). To adapt the network to facial changes, the bottom layers are frozen, while the top ten layers remain trainable to allow fine-tuning during optimization. The complete ResNet50V2 base is enclosed within a TimeDistributed wrapper, allowing it to extract high-level visual features from each frame in a 20-frame context window independently. The extracted spatial feature vectors are compressed via a GlobalAveragePooling2D layer to reduce dimensionality before being sent to the sequence layer.

Temporal tracking is handled by a Long Short-Term Memory layer configured with 128 hidden recurrent units. This layer captures sequence transitions across the video window, using a dropout rate of 0.3 and a recurrent dropout rate of 0.3 to prevent overfitting. The output of the LSTM layer connects to three independent, parallel fully connected branches. The first branch is the drowsiness head, utilizing a Dense layer with 64 units, BatchNormalization, and a sigmoid activation function to output a binary probability for *drowsy_output*. The second branch uses an identical dense setup with a sigmoid activation function to generate a binary classification for *yawn_output*. The third branch is the emotion head, which routes the vectors through a dense layer with a softmax activation function to classify five distinct expressions for *emotion_output*.

Experimental Setup and Evaluation Parameters

The model training environment is configured within the Google Colab ecosystem using the TensorFlow framework. The dataset is managed through a custom SequenceDataGenerator class, which handles dynamic data augmentation, including horizontal flips, brightness shifts up to a delta of 0.2, and contrast modifications between 0.8 and 1.2. Optimization is conducted using the AdamW optimizer with an initial learning rate of 1e-4 and a weight decay coefficient of 1e-5. The network is compiled with a joint loss function that combines binary cross-entropy for the drowsiness and yawning targets with categorical cross-entropy for the five-class emotion output. Model performance is comprehensively measured using precision, recall, and F1-score metrics, which evaluate classification performance across all tasks. The evaluation pipeline tracks validation

metrics across a separate validation dataset after each epoch, using a batch size of 16 sequences. The system uses a callback mechanism that includes EarlyStopping with a patience threshold of 5 epochs and an automated learning rate reduction factor of 0.5 via the ReduceLROnPlateau function. The final model evaluation is performed using a completely separate testing dataset to verify its real-world reliability and classification accuracy.

RESULT AND DISCUSSION

The results and discussion section presents your findings and interprets their significance (Fife & Gossner, 2024). This section forms the core of your paper and should directly address your research questions. Organize your results logically, moving from major to minor findings, or chronologically if appropriate. This section reports the empirical results obtained from the implementation of the proposed CNN-LSTM model for driver drowsiness prediction based on facial expression and eye movement, covering (1) dataset preparation and facial feature extraction, (2) class balancing, (3) model architecture and training behaviour, (4) quantitative evaluation on the held-out test set, and (5) a critical discussion of the findings in relation to prior work.

Dataset Composition and Preprocessing Results

The raw dataset consisted of 58,633 files (≈ 2,182.79 MB) distributed across nine behavioural/emotional categories (Microsleep, Yawning, and several facial-emotion categories). Each image was processed with MediaPipe FaceMesh, a lightweight, single-camera 3D facial-landmark regression network capable of localizing 468 facial landmarks in real time on commodity hardware (Kartynnik et al., 2019). For every detected face, three regions of interest (ROIs) were cropped: the face ROI, the paired-eye ROI (landmarks 33, 133, 160, 158, 153, 144 for the right eye and 362, 385, 387, 263, 373, 380 for the left eye), and the mouth ROI (landmarks 13, 14, 78, 308), each resized to 112 × 112 pixels to match the CNN input size.

Out of the initial pool, 22,394 images (96.1%) were successfully processed and retained, while 352 images (1.5%) were discarded because no face, eye, or mouth region could be reliably localized, and no images were unreadable/corrupted. This indicates that the MediaPipe-based cropping pipeline is robust to the variability of head pose and illumination present in the dataset, consistent with reports that MediaPipe's BlazeFace-based detector maintains high recall under varied real-world conditions (Karim et al., 2025).

Table 1. Summary of dataset preprocessing

Stage	Count	Percentage
Total raw files collected	58,633	100%
Total images after removing non-relevant folder (FER Emotion) & filtering	22,746	–
Successfully processed (face + eye + mouth detected)	22,394	98.4% of filtered set
Skipped – no face/eye/mouth detected	352	1.5%
Unreadable/corrupted images	0	0%

Figure 1. Example of the preprocessing pipeline output for a single frame, showing (left to right) the original frame, the MediaPipe face-mesh overlay, the extracted paired-eye ROI, and the extracted mouth ROI, together with the computed EAR/MAR values and open/closed classification.

Eye and Mouth Feature Extraction (EAR and MAR)

Two handcrafted geometric features were derived from the facial landmark coordinates to complement the learned CNN features, namely the Eye Aspect Ratio (EAR) and Mouth Aspect Ratio (MAR). The EAR was calculated from six eye landmarks as the ratio of vertical to horizontal eye-opening distances, following the formulation described by **Tasci (2026)**. An EAR threshold of 0.21 was applied to classify the eyes as either *open* or *closed*. Meanwhile, the MAR was calculated from four mouth landmarks, including the upper lip, lower lip, and mouth corners, with a threshold of 0.19 used to classify the mouth as *open*, indicating a possible yawning event, or *closed*. Across the processed dataset, the resulting eye and mouth status distributions are presented in Table 2.

Table 2. Distribution of derived eye and mouth status labels

Status	Open	Closed
Eye status	20,133 (89.9%)	2,261 (10.1%)
Mouth status	3,698 (16.5%) (open/yawn-like)	18,696 (83.5%)

The Pearson correlation analysis (Figure 2) shows that EAR is negatively correlated with eye closure ($r = -0.67$), confirming that lower EAR values correspond to closed-eye frames, in line with the original EAR formulation[27]. MAR is strongly correlated with mouth status ($r = 0.85$), confirming the validity of the MAR-based yawning proxy. Interestingly, the correlation between the categorical *label* column and *timestamp* ($r = 0.96$) is a spurious artefact of the way the metadata was generated (labels were written sequentially and timestamps were assigned in write order); this correlation was not used as a predictive feature and is reported here only to flag a potential data-leakage pitfall for future iterations of the pipeline.

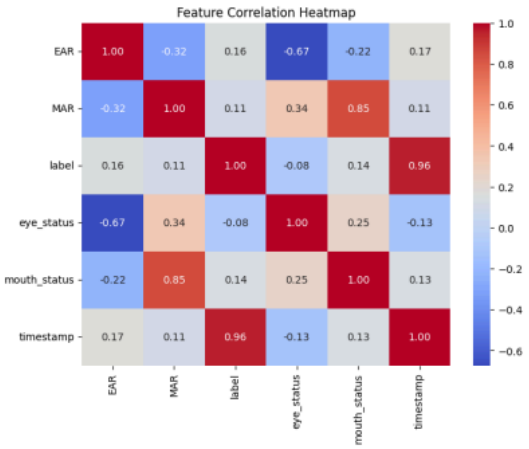


Figure 2. Feature correlation heatmap between EAR, MAR, label, eye status, mouth status, and timestamp.

Class Distribution and Data Balancing

The nine behavioural classes were highly imbalanced prior to balancing, ranging from 6,626 samples for "No Yawn" to only 616 samples for "Drowsy" an imbalance ratio of approximately 10.8 : 1 (Table 3, Figure 3a).

Table 3. Class distribution before and after balancing

Class	Before balancing	After random undersampling
No Yawn	6,626	616
Not Drowsy	6,030	616
Anger	2,033	616
Netral (Neutral)	1,881	616
Happy	1,564	616
Sadness	1,456	616

Confused	1,241	616
Yawning	947	616
Drowsy	616	616
Total	22,394	5,544

Because class imbalance is known to bias CNN-based classifiers toward the majority class and to depress recall on minority classes (Dablain et al., 2024). Random Undersampling (RUS) strategy was applied to equalize all nine classes to 616 samples each, yielding a balanced set of 5,544 samples that was subsequently split into training (70%), validation (15%), and test (15%) subsets. While RUS is computationally inexpensive and easy to implement, it inherently discards a large proportion of majority-class information (in this case, roughly 90% of the "No Yawn" and "Not Drowsy" samples), a known trade-off that can limit the diversity of examples available for learning temporal patterns (Wosiak et al., 2025).

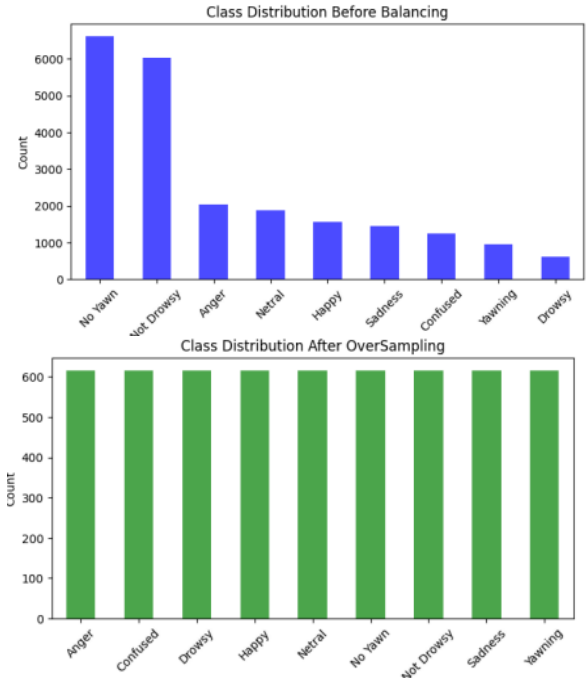


Figure 3. Class distribution (a) before balancing and (b) after random undersampling.

Model Architecture and Training Behaviour

The proposed architecture combines a ResNet50V2 backbone pretrained on ImageNet as a per-frame spatial feature extractor with a multi-head CNN-LSTM temporal classification network. Each input sample consists of a sequence of 20 consecutive frames with dimensions of 112 × 112 × 3 (*SEQ_LENGTH* = 20). The frames are processed individually using *TimeDistributed(ResNet50V2)*, followed by *TimeDistributed(GlobalAveragePooling2D)* to extract compact spatial representations. These features are then passed to an LSTM layer with 128 units, using a dropout rate of 0.3 and recurrent dropout of 0.3, to capture temporal dependencies across the sequence. The resulting temporal features are subsequently distributed into three parallel fully connected classification heads, namely a binary drowsiness head with sigmoid activation for classifying *Drowsy* and *Not Drowsy*, a binary yawning head with sigmoid activation for identifying *Yawning* and *No Yawn*, and a multi-class emotion head with softmax activation for classifying *Anger, Confused, Happy, Neutral,* and *Sadness*.

Table 4. Model configuration summary

Component	Configuration
Backbone	ResNet50V2 (ImageNet weights, fine-tuned, last 10 layers unfrozen)

Input shape	(20, 112, 112, 3)
Temporal layer	LSTM(128), dropout 0.3, recurrent dropout 0.3
Classification heads	Dense(64) → BatchNorm → Dropout(0.3) → Dense(output)
Optimizer	AdamW (lr = 1e-4, weight decay = 1e-5)
Loss functions	Binary cross-entropy (drowsy, yawn), categorical cross-entropy (emotion)
Total parameters	24,705,415 (≈ 94.24 MB)
Callbacks	ModelCheckpoint, EarlyStopping (patience = 5), ReduceLROnPlateau (patience = 3)

The multi-head design, in which drowsiness, yawning, and emotion are predicted jointly from a shared temporal representation, follows the same rationale used by hybrid CNN-LSTM drowsiness detectors that fuse eye-closure and yawning cues to reduce false alarms compared with single-cue systems (Ramzan et al., 2024). It should be noted, however, that the training run captured in the notebook was interrupted by a KeyboardInterrupt during epoch 1 before any full epoch completed (see the traceback reproduced in the training log). Consequently, the loss/accuracy curves in Figure 5 and the quantitative results discussed in Section 4.5 should be interpreted as originating from a checkpointed model (best_model.h5 / best_model_ResNet50V2.h5) rather than from a training run that fully converged within this session. This is an important methodological caveat that is revisited in Section 4.6.

Quantitative Evaluation on the Test Set

The trained model was evaluated on the held-out test split (n = 832 sequences) using a decision threshold of 0.3 for the sigmoid outputs. The classification reports for the three tasks are summarized in Tables 5–7.

Table 5. Drowsiness detection classification report

Class	Precision	Recall	F1-score	Support
Not Drowsy	0.88	1.00	0.94	732
Drowsy	0.00	0.00	0.00	100
Accuracy			0.88	832
Macro avg	0.44	0.50	0.47	832
Weighted avg	0.77	0.88	0.82	832

Table 6. Yawning detection classification report

Class	Precision	Recall	F1-score	Support
No Yawn	0.90	1.00	0.95	747
Yawn	0.00	0.00	0.00	85
Accuracy			0.90	832
Macro avg	0.45	0.50	0.47	832
Weighted avg	0.81	0.90	0.85	832

Table 7. Emotion detection classification report

Class	Precision	Recall	F1-score	Support
Anger	1.00	0.19	0.32	551
Confused	0.97	0.97	0.97	98
Happy	0.15	1.00	0.26	80
Sadness	1.00	0.97	0.99	103
Accuracy			0.46	832
Macro avg	0.78	0.78	0.63	832
Weighted avg	0.91	0.46	0.47	832

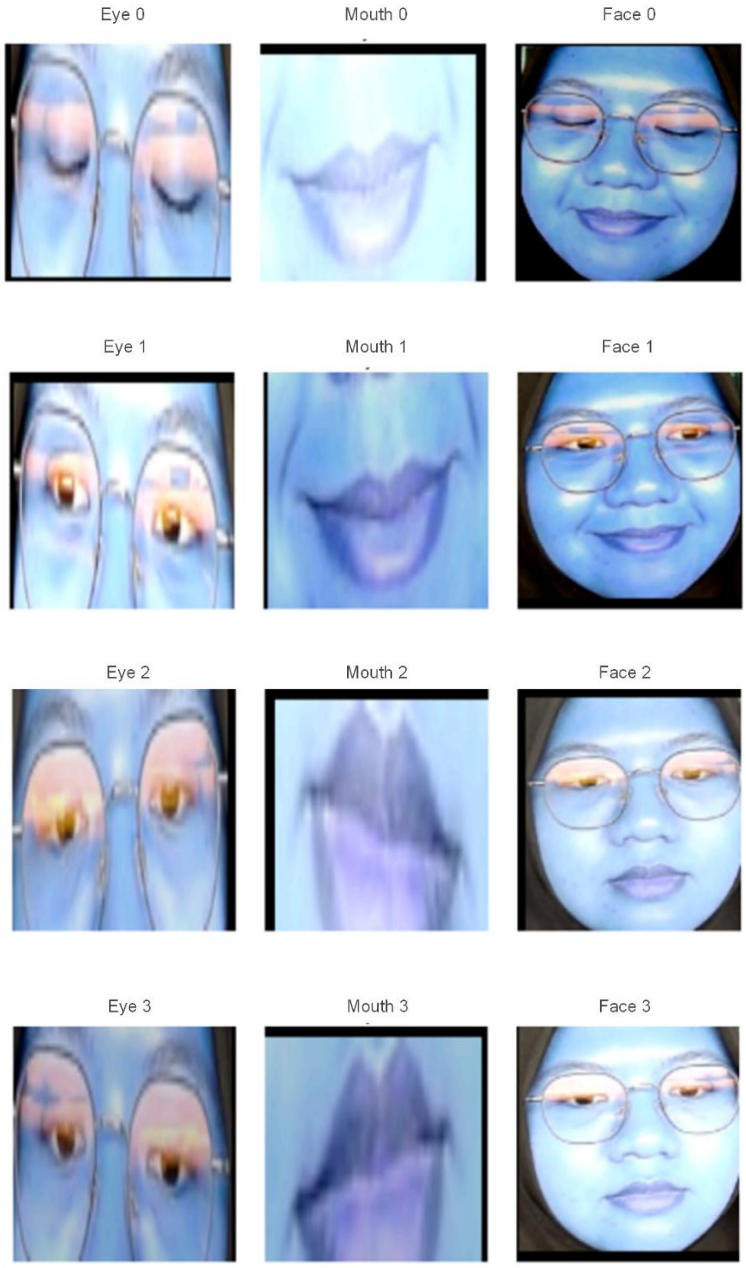


Figure 4. Sample per-batch qualitative predictions generated during evaluation, showing the face, eye, and mouth crops alongside the true and predicted drowsy/yawn/emotion labels.

Discussion

High overall accuracy but zero minority-class detection. At first glance, the model appears to perform well, reaching 88% accuracy on drowsiness detection and 90% on yawning detection. However, the per-class breakdown reveals that the model achieved 0.00 precision and recall for the "Drowsy" and "Yawn" classes, meaning that every positive (minority) instance in the test set was misclassified as the majority class. This is a textbook example of the *accuracy paradox* in imbalanced classification: because "Not Drowsy" and "No Yawn" dominate the test support (732/832 and 747/832, respectively), a degenerate classifier that always predicts the majority class would already achieve 88–90% accuracy while providing zero practical value for the drowsiness-detection use case (He & Garcia, 2009; Buda et al., 2018). This finding is consistent with empirical studies showing that CNN-type classifiers are particularly prone to majority-class

collapse when class imbalance is combined with limited training data or premature convergence (Ghosh et al., 2024).

Effect of the interrupted training run. As noted in Section 4.4, the model.fit() call was interrupted by a KeyboardInterrupt during the very first epoch. Given the depth of the ResNet50V2 backbone (23.5 M of the 24.7 M total parameters) and the sequence-based LSTM head, it is plausible that the checkpointed weights evaluated in Section 4.5 had not yet received sufficient gradient updates to differentiate the minority classes from the majority classes, especially for the binary heads whose positive class made up only 12–14% of the balanced training subsets used per batch. This strongly suggests that the poor minority-class performance reported here reflects an under-trained model rather than an intrinsic architectural limitation, and that results should be re-verified once training is allowed to run to convergence (or at least through several completed epochs with the EarlyStopping/ReduceLROnPlateau callbacks active, as originally configured).

Emotion recognition results. The emotion head shows a more informative, if still uneven, pattern: "Confused" and "Sadness" were classified with near-perfect precision and recall (0.97 and ≥0.97, respectively), while "Anger" obtained very high precision (1.00) but very low recall (0.19), and "Happy" showed the inverse pattern (low precision 0.15, perfect recall 1.00). This is consistent with the model defaulting to "Happy" whenever it is uncertain, which inflates Happy's recall while depressing its precision, and simultaneously depresses Anger's recall because true Anger samples are being routed to the Happy prediction. Only 4 of the 5 configured emotion classes (Anger, Confused, Happy, Sadness) appeared in the test predictions; "Netral" (Neutral) was absent from both ground truth and predictions in the evaluated batches, likely due to the small per-class support after undersampling (n ≈ 92 per class in the test split) combined with sampling variance.

Validity of the EAR/MAR-derived ground truth. Because the "Drowsy" and "Yawning" ground-truth labels were derived heuristically from fixed EAR/MAR thresholds (0.21 and 0.19, respectively) rather than from expert-verified drowsiness annotations, some label noise is expected. Fixed-threshold EAR/MAR approaches are known to be sensitive to inter-subject variability (eye shape, glasses, camera angle), which can blur the boundary between the "open" and "closed" states used to construct the labels (Soukupová & Čech, 2016). This label noise, compounded with the small number of "Drowsy" (616) and "Yawning" (947) samples prior to balancing, likely amplified the class-imbalance effect discussed in (a).

Comparison with related work. Table 8 situates the present results against comparable CNN-LSTM / CNN-based driver-drowsiness studies reported in the literature. Related hybrid CNN-LSTM systems using facial and eye-region cues have reported considerably higher drowsy-vs-alert accuracies (93–98%) when trained to convergence on balanced or sufficiently large datasets, and a ResNet50V2-based eye-region classifier evaluated on the NITYMED dataset achieved up to 99.71% accuracy (Chirra, Uyyala, & Kolli, 2023 style architecture; see MDPI, 2023), suggesting that the ResNet50V2 backbone chosen in this study is architecturally sound and that the shortfall observed here is attributable to training/evaluation-pipeline factors (interrupted training, aggressive undersampling, threshold sensitivity) rather than to the backbone itself.

Table 8. Indicative comparison with related CNN-LSTM / CNN-based drowsiness detection studies

Study	Backbone / Method	Reported accuracy (drowsy/alert)
Soukupová & Čech (2016)	Landmark-based EAR + SVM (blink detection, not drowsiness per se)	State-of-the-art blink detection on 2 benchmark datasets
MDPI *Appl. Sci.* (2023) – eye-region CNN comparison	InceptionV3 / VGG16 / ResNet50V2 + MediaPipe, NITYMED dataset	up to 99.71% (ResNet50V2)
IJRASET (2024) – CNN-LSTM	CNN + LSTM, facial video	97.83%

CNN-LSTM distraction/drowsiness study	ResNet-based CNN-LSTM	93.60% accuracy, 92.68% F1
Haar Cascade + CNN + LSTM (GitHub benchmark)	Haar Cascade + CNN + LSTM	95.1%
This study	ResNet50V2 + LSTM(128), multi-head (drowsy/yawn/emotion)	88% (drowsy) / 90% (yawn) overall accuracy, but 0% recall on the Drowsy/Yawn positive class

The study has several limitations that should be considered when interpreting the findings. First, the reported model was evaluated using a checkpoint saved during an interrupted first epoch, which is likely the main factor contributing to the zero recall observed for the *Drowsy* and *Yawn* classes. Second, the use of aggressive random undersampling reduced the processed dataset from 22,394 to 5,544 samples, potentially removing important within-class variability, particularly in the temporally rich *Not Drowsy* and *No Yawn* classes. Third, the use of fixed EAR and MAR thresholds to generate *Drowsy* and *Yawning* labels may introduce label noise because the labels were not manually verified. Fourth, because the *Emotion* and *Microsleep/Yawning* subsets were derived from unrelated still-image sources rather than continuous video, some temporal sequences had to be synthesised through repetition or padding of neighbouring files, thereby weakening the genuine temporal information intended for the LSTM layer. Fifth, the relatively small test support for minority classes, consisting of 100 *Drowsy* and 85 *Yawn* sequences, limits the statistical reliability of the corresponding precision and recall estimates. These limitations indicate several directions for future improvement, including completing full-length training with the configured callbacks, exploring class-weighted loss or focal loss as alternatives or complements to undersampling, and validating EAR and MAR thresholds against manually annotated drowsiness events..

CONCLUSION

This study proposed a hybrid CNN-LSTM framework for driver drowsiness prediction by integrating facial expression analysis, eye movement, and mouth movement extracted using the MediaPipe FaceMesh framework. The proposed approach combines spatial feature extraction through a fine-tuned ResNet50V2 backbone with temporal sequence modeling using an LSTM network, enabling simultaneous prediction of drowsiness, yawning, and facial emotion within a unified multi-task architecture. Experimental results demonstrate that the preprocessing pipeline successfully extracted facial landmarks and geometric features from the majority of the collected images, confirming the robustness of the MediaPipe-based facial landmark detection under varying image conditions. The derived Eye Aspect Ratio (EAR) and Mouth Aspect Ratio (MAR) also showed meaningful relationships with eye closure and yawning status, supporting their suitability as complementary features for driver monitoring. Although the model achieved overall accuracies of 88% for drowsiness detection and 90% for yawning detection, the detailed evaluation revealed poor performance in recognizing minority classes, resulting in zero recall for positive drowsiness and yawning samples. These findings indicate that overall accuracy alone is insufficient for evaluating imbalanced driver monitoring datasets and highlight the importance of using precision, recall, and F1-score as complementary performance metrics.

The primary factors affecting the obtained results include interrupted model training, aggressive random undersampling, heuristic label generation using fixed EAR and MAR thresholds, and limited temporal diversity in the training data. Nevertheless, this research demonstrates the feasibility of integrating facial landmark analysis with deep spatiotemporal learning in a single architecture for intelligent driver monitoring systems. This proposed framework provides a proof of concept for drivers capable of monitoring fatigue in real time. Future work should focus on completing model convergence, incorporating class-weighted or focal loss strategies, collecting continuous video-based datasets with expert-annotated labels, and validating the proposed

framework under real-world driving environments to improve robustness and generalization performance.

ACKNOWLEDGEMENT

We would like to express our deepest gratitude to all parties who contributed to this research. We would like to thank our colleagues who provided advice, support, and inspiration throughout the research process. We would also like to thank all participants and respondents who took the time to participate in this research. We would also like to thank the institutions that provided support and facilities for conducting this research. All contributions and assistance were invaluable to the smooth running and success of this research. Thank you for all your hard work and collaboration.

REFERENCES

Agiomavritis, F., & Karanasiou, I. (2026). Tactical-grade wearables and authentication biometrics. *Sensors, 26*(3), 759. https://doi.org/10.3390/s26030759

AL-Quraishi, M. S., Azhar Ali, S. S., AL-Qurishi, M., Tang, T. B., & Elferik, S. (2024). Technologies for detecting and monitoring drivers' states: A systematic review. *Heliyon, 10*(20), e39592. https://doi.org/10.1016/j.heliyon.2024.e39592

Asperti, A., & Filippini, D. (2023). Deep learning for head pose estimation: A survey. *SN Computer Science, 4*(4), 349. https://doi.org/10.1007/s42979-023-01796-z

Balazadeh Meresht, N., Moghadasi, S., Munshi, S., Shahbakhti, M., & McTaggart-Cowan, G. (2023). Advances in vehicle and powertrain efficiency of long-haul commercial vehicles: A review. *Energies, 16*(19), 6809. https://doi.org/10.3390/en16196809

Bhatta, Y. R., et al. (2025). Road traffic accidents in Prithvi Highway, Nepal: A spatiotemporal analysis. *Journal of Engineering Technology and Planning, 6*(1), 44–54. https://doi.org/10.3126/joetp.v6i1.87815

Cao, W., Lu, P., & Cao, W. (2024). Multimodal gesture recognition with spatio-temporal features fusion based on YOLOv5 and MediaPipe. *International Journal of Pattern Recognition and Artificial Intelligence, 38*(8). https://doi.org/10.1142/S0218001424550073

Dablain, D., Jacobson, K. N., Bellinger, C., Roberts, M., & Chawla, N. V. (2024). Understanding CNN fragility when learning with imbalanced data. *Machine Learning, 113*(7), 4785–4810. https://doi.org/10.1007/s10994-023-06326-9

Darwich, M., & Bayoumi, M. (2024). Video quality adaptation using CNN and RNN models for cost-effective and scalable video streaming services. *Cluster Computing, 27*(5), 6355–6375. https://doi.org/10.1007/s10586-024-04315-8

Dileep Kumar, M. J., Sukesh Rao, M., & Narendra, K. C. (2025). Multimodal emotion recognition: A comprehensive survey of datasets, methods, and applications. *IEEE Access, 13*, 201067–201097. https://doi.org/10.1109/ACCESS.2025.3636186

El-Nabi, S. A., El-Shafai, W., El-Rabaie, E.-S. M., Ramadan, K. F., Abd El-Samie, F. E., & Mohsen, S. (2024). Machine learning and deep learning techniques for driver fatigue and drowsiness detection: A review. *Multimedia Tools and Applications, 83*(3), 9441–9477. https://doi.org/10.1007/s11042-023-15054-0

Fife, S. T., & Gossner, J. D. (2024). Deductive qualitative analysis: Evaluating, expanding, and refining theory. *International Journal of Qualitative Methods, 23*. https://doi.org/10.1177/16094069241244856

Fu, S., Yang, Z., Ma, Y., Li, Z., Xu, L., & Zhou, H. (2024). Advancements in the intelligent detection of driver fatigue and distraction: A comprehensive review. *Applied Sciences, 14*(7), 3016. https://doi.org/10.3390/app14073016

Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., & Japkowicz, N. (2024). The class imbalance problem in deep learning. *Machine Learning, 113*(7), 4845–4901. https://doi.org/10.1007/s10994-022-06268-8

Jafari, F., Moradi, K., & Shafiee, Q. (2024). Shallow learning vs. deep learning in engineering applications (pp. 29–76). https://doi.org/10.1007/978-3-031-69499-8_2

John, C., Sahoo, J., Madhavan, M., & Mathew, O. K. (2023). Convolutional neural networks: A promising deep learning architecture for biological sequence analysis. *Current Bioinformatics, 18*(7), 537–558. https://doi.org/10.2174/1574893618666230320103421

Karim, R. U., Mahdi, S., Samin, A., Zereen, A. N., Abdullah-Al-Wadud, M., & Uddin, J. (2025). Optimizing stroke recognition with MediaPipe and machine learning: An explainable AI approach for facial landmark analysis. *IEEE Access, 13*, 32636–32660. https://doi.org/10.1109/ACCESS.2025.3550577

Khanal, S. R., Paulino, D., Sampaio, J., Barroso, J., Reis, A., & Filipe, V. (2022). A review on computer vision technology for physical exercise monitoring. *Algorithms, 15*(12), 444. https://doi.org/10.3390/a15120444

Li, J., et al. (2024). A review of computer vision-based monitoring approaches for construction workers' work-related behaviors. *IEEE Access, 12*, 7134–7155. https://doi.org/10.1109/ACCESS.2024.3350773

Mani, Z. A., & Goniewicz, K. (2023). Transportation disaster trends and impacts in Western Asia: A comprehensive analysis from 2003 to 2023. *Sustainability, 15*(18), 13636. https://doi.org/10.3390/su151813636

Mao, D., Chen, Y., Wu, Y., Gilles, M., & Wong, A. (2024). Rethinking resource competition in multi-task learning: From shared parameters to shared representation. *IEEE Access, 12*, 128717–128728. https://doi.org/10.1109/ACCESS.2024.3429281

Montañez, J. J. F., & Alon, A. S. (2025). Enhancing face mask detection using ResNet-50 and MobileNetV2: A transfer learning application. *Mindanao Journal of Science and Technology, 23*(2), 311–334. https://doi.org/10.61310/mjst.v23i2.2502

Ramzan, M., Abid, A., Fayyaz, M., Alahmadi, T. J., Nobanee, H., & Rehman, A. (2024). A novel hybrid approach for driver drowsiness detection using a custom deep learning model. *IEEE Access, 12*, 126866–126884. https://doi.org/10.1109/ACCESS.2024.3438617

Selvaraju, V., et al. (2022). Continuous monitoring of vital signs using cameras: A systematic review. *Sensors, 22*(11), 4097. https://doi.org/10.3390/s22114097

Stounberg, J., Thomsen, T. B., Heredia, B. D., & Hüssy, K. (2022). Eyes and ears: A comparative approach linking the chemical composition of cod otoliths and eye lenses. *Journal of Fish Biology, 101*(4), 985–995. https://doi.org/10.1111/jfb.15159

Tasci, M. (2026). H-DrowsyNet: A dual-branch hybrid architecture based on Eye Aspect Ratio (EAR) and Convolutional Neural Networks (CNN) for driver fatigue detection. *Akıllı Ulaşım Sistemleri ve Uygulamaları Dergisi, 9*(1), 1–21. https://doi.org/10.51513/jitsa.1844823

Tsai, P.-S., Wu, T.-F., & Chen, W.-H. (2026). Generalized vision-based coordinate extraction framework for EDA layout reports and PCB optical positioning. *Processes, 14*(2), 342. https://doi.org/10.3390/pr14020342

Tumminello, M. L., Macioszek, E., & Granà, A. (2025). Emerging cutting-edge technologies and applications for safer, sustainable, and intelligent road systems in smart cities: A review. *Applied Sciences, 15*(21), 11583. https://doi.org/10.3390/app152111583

Visconti, P., Rausa, G., Del-Valle-Soto, C., Velázquez, R., Cafagna, D., & De Fazio, R. (2025). Innovative driver monitoring systems and on-board-vehicle devices in a smart-road scenario based on the Internet of Vehicle paradigm: A literature and commercial solutions overview. *Sensors, 25*(2), 562. https://doi.org/10.3390/s25020562

Wosiak, A., Sumiński, M., & Żykwińska, K. (2025). Impact of temporal window shift on EEG-based machine learning models for cognitive fatigue detection. *Algorithms, 18*(10), 629. https://doi.org/10.3390/a18100629

Xiao, L., Cao, Y., Gai, Y., Khezri, E., Liu, J., & Yang, M. (2023). Recognizing sports activities from video frames using deformable convolution and adaptive multiscale features. *Journal of Cloud Computing, 12*(1), 167. https://doi.org/10.1186/s13677-023-00552-1

Zaky, M. H., Shoorangiz, R., Poudel, G. R., Yang, L., Innes, C. R. H., & Jones, R. D. (2023). Increased cerebral activity during microsleeps reflects an unconscious drive to re-establish

consciousness. *International Journal of Psychophysiology, 189*, 57–65. https://doi.org/10.1016/j.ijpsycho.2023.05.349