VI. International Conference on Electrical, Computer and Energy Technologies (ICECET 2026)
6-9 July 2026, Rome-Italy

Baseline vs Exploratory Deep Learning Video Models for Driver Drowsiness Detection: A Multi-Dataset Evaluation

Abhijith Pradeep
School of Engineering
University of Warwick
Coventry, United Kingdom
abhijith.pradeep@outlook.com

Wael AlSaafin
Warwick Manufacturing Group
University of Warwick
Coventry, United Kingdom
Wael.Alsaafin@warwick.ac.uk

***Abstract*—Driver drowsiness detection is challenging to implement in the real world due to various shifts in the user's environment i.e. illumination, eyewear and head pose. This study intends to evaluate the capability of the current industry standards which are achieved with the baseline models (PERCLOS & Rule-Based Fusion of the eyelid and mouth dynamics) which are thus compared with exploratory deep-learning models (CNN-GRU, 3D-CNN and Temporal Shift Modelling). The uniqueness of this study is that all experiments are conducted with subject-independent splits paired with strict processing of the datasets (112×112, 10 fps and 16 frames). This is crucial when we are training, testing and validating models from three globally recognised public datasets (UTA-RLDD, YawDD and DROZY). The results are tested and evaluated using the F1-score, PR-AUC and other metrics under a unified protocol. Cross-domain robustness is explicitly quantified by measuring performance degradation between internal and external test sets, enabling analysis of domain shifts. Evaluation shows that deep models definitely outperform baseline models. R(2+1)D-18 achieved the best internal testing (UTA-RLDD) F1-score, while CNN-GRU had the best performance with external testing (DROZY) showing significant robustness. Currently based on the results, CNN-GRU is the model which would be most apt to be implemented in the real-world due to its performance and low computational cost. The results help to visualise the reactions of the models to varying domain shifts.**

***Index Terms*—driver monitoring, driver drowsiness detection, PERCLOS, facial landmarks, deep-learning, temporal modeling, domain shift, subject independent evaluation, benchmarking, UTA-RLDD**

I. Introduction

Throughout the years, drowsy driving has remained a major cause of accidents. According to the National Highway Traffic Safety Administration (NHTSA), 17.6% of fatal crashes from 2017–2021 involved a drowsy driver, resulting in an estimated 29,834 deaths over five years [1]. Additionally, in 2022, around 1300 collisions in the UK resulted in injuries [2]. These statistics highlight the need for reliable driver monitoring within advanced driver assistance systems (ADAS) to improve road safety [3]. This abundance of accidents can be reduced with the help of driver monitoring which is the ability to pick up the driver's facial cues such as eye closure, yawning and slow blinking. The existing systems focus on a single indicator of analysis (e.g. blink duration or eyelid closure percentage) while this is effective, it is vastly inaccurate due to the various edge cases involved such as brightness changes, head poses and eyewear on the driver [4].

This study establishes a unified benchmarking framework for comparing baseline and deep-learning models. A subject-independent data split, along with consistent pre-processing and evaluation protocols, is applied across all datasets to prevent data leakage and ensure unbiased evaluation.

The model is trained and tested on three different globally recognised datasets (UTA-RLDD, YawDD and DROZY) [5] [6] [7]. All these datasets account for the different domain shifts involved in driver monitoring and the environment. Percentage of closed eyes (PERCLOS) is widely used as a baseline in the industry and Rule-based Fusion (RBF) is a combination of both PERCLOS and the mouth aspect ratio (MAR) to detect yawning as well [4] [8]. These are the baseline models which are evaluated. As for the deep-learning models, it involves Temporal Shift Modelling (TSM), CNN-GRU, R(2+1)D-18 (3D CNN) [9] [10] [11]. These are analysed for accuracy, computational efficiency and robustness in detecting drowsiness and hence compared with the baseline models under the same protocol. The performance of the models is analysed using F1-score and PR-AUC with the drowsiness thresholds tuned only on validation data [12].

Main contributions of this study:

1) A unified multi-dataset benchmarking protocol for driver drowsiness across UTA-RLDD, YawDD and DROZY
2) Subject-independent splitting to prevent data leakage across training, testing and validation
3) Comparison of industry baselines against temporal deep video models under identical evaluation protocol

4) Cross-domain robustness measured using F1-Score across internal and external tests
5) Analysis of robustness, computational costs and accuracy for real-world deployment

The rest of this paper is organised as follows. Related works are addressed in Section II, the problem is defined in Section III, pre-processing works and model specifics are discussed in Sections IV and V. Experimental analysis is presented in Section VI, followed by results analysis in Section VII. Finally, discussions and conclusions are derived in Sections VIII and IX.

II. Related Works

A. Classical Vision-Based Approaches

The existing systems are mainly based on the analysis of PERCLOS and this is set as the key industrial baseline [4]. Although eyelid closure is the most determining factor in detecting drowsiness, there are various limitations in the system involved with lighting, eyewear, camera propagation and head poses which could lead to false positives and negatives in driver drowsiness detection [13]. On the other hand, the RBF model combines the analysis of PERCLOS and MAR to identify whether the user is drowsy [8]. This improves the robustness of the system as it is analysing both eye closure and mouth gape to detect drowsiness.

B. Deep Learning Video Models

Deep-learning models are very useful as they can analyse the temporal aspect as well as the spatial features of the video to specifically capture better dynamics and results than static image analysis [14]. All models aim to visualise how facial features move over time. CNN combined with GRU helps identifying facial features and noticing their changes over time, whereas 3D-CNN analyses spatiotemporal features of the user and requires significant computational power [9] [10]. The TSM is a lightweight model which aims to do the same with lower computational power compared to the other two models [15].

C. Evaluation and Testing Protocol

These models are complex and produce different results if there are inconsistent methodologies and evaluation protocols undergone in the testing and training phase. This study aims to implement and train under a unified protocol with subject-independent testing to ensure the evaluation metrics such as F1-score and PR-AUC accurately represent the model performance [12] [16]. Additionally, we focused on the performance of the models under domain shifts and attempted to improve the robustness by using multiple datasets which represent the said domain shifts.

III. Problem Definition and Datasets

A. Binary Video Classification

This study attempts to detect driver drowsiness as a binary video classification task. Each input clip per subject is classified as a label of 0 (not drowsy) or 1 (drowsy). A binary setup is very efficient as there are no other median classifications involved and it is very direct in detecting drowsiness which is apt for a vehicular safety system [17].

B. Datasets

For this study, the models were trained with three different universally recognised datasets: UTA-RLDD, YawDD and DROZY.

UTA-RLDD is the primary dataset for our study and sets the baseline for how drowsiness is depicted in humans. This dataset contains clips in natural lighting of subjects in three different levels: Alert, Low-Vigilance and Drowsiness [5].

YawDD is a secondary dataset involved in training as it allows us to identify yawning within driver monitoring [6]. Yawning is commonly associated with drowsiness and YawDD has provided data for subjects who are alert, talking and yawning [13]. This allows the model to understand the difference between talking and yawning, hence creating a robust model.

Lastly, DROZY is a dataset which is used strictly for testing drowsiness under low-light conditions and ensuring the robustness of the model. This dataset is unique as it contains electroencephalogram signals of the subjects so this allows classification to truly depict how facial features change with drowsiness [7].

Table I shows the different domain shifts and features accounted for by the multiple datasets present. There were a total of 116 subjects involved in this study. The models are trained, tested and validated on approximately 76751 clips across the 3 datasets.

TABLE I
Dataset Spread of Domain Shifts

Feature	**UTA-RLDD**	**YawDD**	**DROZY**
Total Subjects	60	42	14
Males	51	21	3
Females	9	21	11
Different Ethnicities	5	4	NA
Glasses/Sunglasses	7	24	0
Facial Hair	24	10	1
Lighting	Indoor	Sunny/Rainy	Low Light
Environment	In Rooms	Lab	Driving

C. Label Mapping

Mapping is how we determine what stages represent the binary video classification of 0 & 1 for drowsiness. With the labelling, it was strict to ensure that the model was accurate in detecting drowsiness and did not misrecognize the data.

Table II shows how each dataset was evaluated and labelled according to binary video classification. UTA-RLDD has the original low vigilance label (Stage 5) labelled 0 as well to ensure that drowsiness detection was completely accurate and did not have false positives. Stage 0 and Stage 10 in UTA-RLDD were for alertness and drowsiness respectively.

TABLE II
DATASET BINARY LABEL METHODOLOGY

Dataset	Original Labels	Mapping (0=Alert, 1=Drowsy)
UTA-RLDD	Stage 0, 5, 10	Stage 0 & 5 $\rightarrow$ 0; Stage 10 $\rightarrow$ 1
YawDD	Normal, Talking, Yawning	Normal & Talking $\rightarrow$ 0; Yawning $\rightarrow$ 1
DROZY	Alert, Drowsy	Alert $\rightarrow$ 0; Drowsy $\rightarrow$ 1

YawDD has only yawning labelled as 1 as it is an indicator of drowsiness and talking labelled as 0 so it would train the model to identify between each. DROZY was labelled using a Karolinska Sleepiness Scale (KSS) which was provided with the EEG signals of the subjects to depict the drowsiness of users. So in this case any subject who had 6 or more in the KSS in the clip was labelled a 1 [18].

D. Subject Split

All subject splits ensure that no subjects appear across training, validation and test sets and all clips for a subject are kept within a single split. This confirms that the model does not generalise on a person's appearance or the environment in the clip when detecting drowsiness. This split is randomly compiled using a script to prevent human bias.

IV. PRE-PROCESSING OF THE CLIPS

All datasets had different video clips of varying lengths. For the model inputs, the clips needed to be of the same format for the uniformity of the testing process.

Each video is clipped at 112×112 format with 16 frames and at 10 fps. The models are trained in this format as this was the minimum quality estimated for CNN networks [19]. Additionally, a stride of 8 frames is used to produce clips that overlap by 50% ensuring that there is increased coverage of the data. This setup not only ensured each frame of each video was captured, it was also computationally feasible to detect drowsiness. All the particular training, testing and validation clips were tracked using manifests to ensure reproducibility of the list of clips and that there was no data leakage. The training and testing of the deep-learning models were executed in Google Colab using an NVIDIA Tesla T4 GPU (12GB VRAM).

V. MODEL SPECIFICS

A. Baseline Methods

Two baseline methods were tested with the help of the MediaPipe FaceLandmarker to identify 468 independent points on the face [20]. The features of the face were extracted using this and were computed on a small interval of frames per clip (4 frames for UTA-RLDD/YawDD and 2 frames for DROZY). The following are the two crucial formulas for the baseline models.

$$\text{EAR} = \frac{\|\mathbf{p}_2 - \mathbf{p}_6\| + \|\mathbf{p}_3 - \mathbf{p}_5\|}{2\,\|\mathbf{p}_1 - \mathbf{p}_4\|} \tag{1}$$

$$\text{MAR} = \frac{\|\mathbf{p}_{\text{upper}} - \mathbf{p}_{\text{lower}}\|}{\|\mathbf{p}_{\text{left}} - \mathbf{p}_{\text{right}}\|} \tag{2}$$

$\|\mathbf{p}_1 - \mathbf{p}_4\|$ defines the maximum horizontal distance in the eye opening while $\|\mathbf{p}_2 - \mathbf{p}_6\|$ and $\|\mathbf{p}_3 - \mathbf{p}_5\|$ represent the maximum vertical distance. The eye closure was approximated using Eye Aspect Ratio (EAR) and yawning was identified with the Mouth Aspect Ratio (MAR). The eye was considered closed if the EAR was less than 0.21 and as yawning if the MAR was above 0.65 [8]. PERCLOS was calculated as the proportion of frames in a clip that had closed eyes.

$$\text{PERCLOS} = \frac{\text{Number of frames with eyes closed}}{\text{Total frames analysed}} \tag{3}$$

The RBF model was more complex compared to PERCLOS as it used a combination of eye and mouth detection [8]. This was done by weighting PERCLOS, the proportion of yawning frames and maximum observed MAR.

$$\text{RBF} = 0.6 \times \text{PERCLOS} + 0.25 \times \text{Yawning Prop.} + 0.15 \times \text{MAR}_{\text{max}} \tag{4}$$

The weighting for each aspect in (4) was decided after tuning on validation data to achieve the maximum F1 score [8]. The final thresholds for PERCLOS (3) and RBF (4) were 0.05 and 0.06 for drowsiness respectively.

B. Deep-Learning Models

Three deep video architectures were tested in this study for their accuracy, robustness and computational cost under a unified protocol. All models were trained on identical clips and output a probability for the drowsy class.

TSM model has a lightweight modelling process which identifies features and tracks their movements across frames. This is vital as TSM attempts to do this without any 3D convolutions. This model was tested to identify the efficiency of lightweight deep models as they would be computationally better for implementation in the real-world [11].

The CNN-GRU model extracts facial features using its CNN backbone and with the assistance of the GRU, it models their movement over time. This allows the model to identify critical drowsiness tellers such as slow blinking, long periods of eye closure and yawning [9]. This model is tested to measure the capability of a sequential temporal model with moderate computational cost.

R(2+1)D-18 (3D CNN) is a deep-learning model that quantifies facial features in 2D and time in 1D components, hence the name [21]. This model identifies and studies spatiotemporal patterns with 3D convolution. This architecture has high motion sensitivity and can learn the movements of the facial features within the convolutional layers. While this model is

very receptive and accurate, it requires a higher computational cost compared to the rest. This model was included in this study to test whether a strong modelling architecture is still as strong across the various domain shifts (Table I) involved [10].

VI. Experimental and Evaluation Method

A. Testing and Threshold Calculation

The most important part of the evaluation protocol was that there was no overlap between the training, testing and validation subjects and it was subject-independent splits across all the datasets. The threshold optimisation and calculation were completed post training and the testing split was to purely test the accuracy of each model. The decision threshold was selected by testing which value produces the highest F1-score and then this locked threshold (τ) value was used for all the test evaluations [12] [16]. This ensures unbiased results through the multiple domain shifts.

Binary Decision Rule:

$$\hat{y} = \mathbf{1}[P(y = 1 \mid x) \geq \tau] \tag{5}$$

The thresholds were used with the binary decision rule to detect drowsiness [17]. These thresholds were difficult to determine as the metrics vary due to the multiple datasets and domain shifts (Table I) so a dual threshold experiment was considered for the deep-learning models [22]. The criteria were to ensure precision was greater than 0.6 with said threshold and the model was not able to achieve this in the threshold tuning with only the validation dataset. Hence, the model (due to the dual threshold) switched to achieve a recall greater than 0.7 [23]. This was set as F1-score depends on precision and recall and setting recall greater than 0.7 results in the best F1 value.

TABLE III
Thresholds Set for Each Model

Model	Threshold (τ)
PERCLOS	0.05
Rule-Based Fusion	0.06
TSM	0.010
CNN-GRU	0.001
R(2+1)D-18	0.004

B. Test Datasets

Individual thresholds (τ) were used with the binary decision rule (5) for each model (Table III) to detect drowsiness.

TABLE IV
Test Dataset Breakdown

Test Dataset	Role	Clips Used
UTA-RLDD	Internal Benchmark	12,078
YawDD	Driving Generalisation	533
DROZY	External Robustness	26,509

Table IV is important as it shows how each dataset is crucial to test the robustness of each model. The UTA dataset is important as it tests the accuracy of the models and sets an internal benchmark. YawDD is the only dataset where there is a driving environment so it tests the capability of the models in this scenario [6]. DROZY is a purely testing dataset which contains low-light footage of the user's drowsiness [7]. There was a total of 39123 clips in the testing data.

C. Metrics

For each model and dataset pair, the performance was evaluated with a variety of metrics. These were the F1-score, precision, recall, specificity, accuracy, PR-AUC and the confusion matrix [16].

F1-score was selected as the primary metric as it is a combination of precision and recall, hence being a valuable measure for the capability of each model. This is because precision is essentially the false positive behaviour whereas recall is the sensitivity to drowsiness which are both important features of a system. On the other hand, specificity tests the capability of the model to detect non-drowsiness. PR-AUC was included as a threshold independent metric that evaluates the quality in the event the threshold is mis-calibrated or skewed [12]. The Cross-domain robustness is calculated using

$$\Delta F1 = F1_{\text{UTA test}} - F1_{\text{DROZY test}} \tag{6}$$

This measures the performance drop between the internal benchmark and the external dataset. This helps understand the robustness of the model [24].

VII. Results

A. Baseline Results

Table V shows the effectiveness of PERCLOS and RBF and it is exactly how the theory expects it to be. Both models have exact same result for UTA & DROZY where yawning is not guaranteed but in YawDD where yawning is present, it significantly improves its performance. The high performance in the UTA dataset is due to the high number of eye closure related signals involved in this dataset. Whereas in DROZY, the performance drops minutely displaying that the models are not robust to domain shifts.

TABLE V
F1-Score of the Baseline Models on All Datasets

Model	UTA F1	YawDD F1	DROZY F1
PERCLOS	0.60	0.23	0.42
Rule-Based Fusion	0.60	0.47	0.42

B. Deep-Learning Video Models Results

Table VI shows the results for the major metrics, F1 and PR-AUC. This clearly conveys that each model is robust enough to identify drowsiness in different domain shifts. The $\Delta F1$ (6) is negative in all the models which supports their robustness. The higher performance on DROZY compared to UTA-RLDD

TABLE VI
DEEP MODELS PERFORMANCE ON INTERNAL AND EXTERNAL TESTING

Model	UTA F1	UTA PR-AUC	DROZY F1	DROZY PR-AUC
TSM	0.605	0.581	0.675	0.623
CNN-GRU	0.622	0.714	0.714	0.695
R(2+1)D-18	0.651	0.510	0.704	0.593

reflects the threshold miscalibration on the internal set rather than cross-domain generalization. Fig. 1 shows the precision-recall curve on the external low-light test conducted. CNN-GRU performed significantly in this displaying robustness to domain shifts with a PR-AUC value of 0.695.

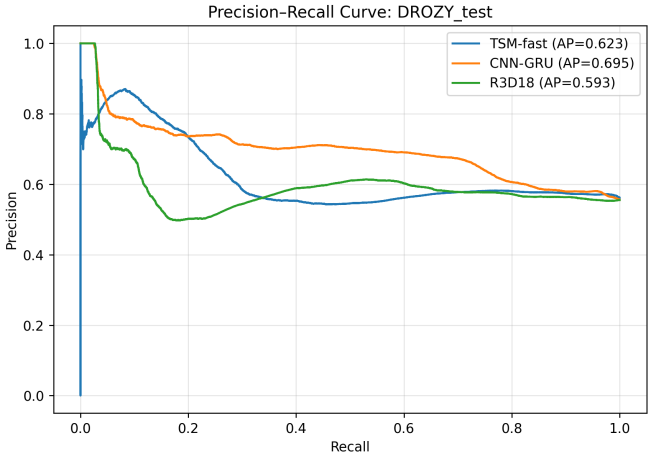

Fig. 1. Precision-Recall Curve of all three deep models on DROZY

R(2+1)D-18 achieved the highest F1-score, indicating its capability within the primary dataset (UTA-RLDD) but CNN-GRU was constantly capable of detecting drowsiness in both datasets and is technically more robust. Additionally, CNN-GRU has a lower computational cost compared to the 3D CNN model. TSM can detect drowsiness somewhat accurately and is robust to domain shifts but is always trailing the other models in performance. On the other hand, YawDD achieved from 0.391 to 0.447 for the F1-score across all 3 models and this was a stark drop compared to the other datasets.

VIII. DISCUSSION

A. Baselines vs Deep Models

PERCLOS and RBF both provide straightforward and effective drowsiness detection methods but are significantly less efficient in cross-domain testing therefore showing a lack of robustness. This is seen particularly in the dataset, DROZY where the F1 score drops to 0.42. RBF had improved performance on YawDD where yawning is the most prominent feature but once again did not show much robustness to the model as seen in Table V. The deep-learning models overall outperform the baselines and show strong robustness to the dataset DROZY (Table VI).

B. Deep-Learning and YawDD Difficulty

When testing for YawDD all the deep video models tended to drop in performance significantly. Upon Grad-CAM analysis, this shift in performance was found to be due to the talking and yawning criteria of YawDD (Table II) [6] [25]. Experiments show that this was inefficient in detecting drowsiness but only yawning and talking. This helps us identify another valid edge case for which the model can be further trained.

C. Threshold Sensitivity & Miscalibration

With the threshold tuning for each model, a crucial part to notice was that τ was awfully low for CNN-GRU with a value of 0.001. This was not a methodological error as the same protocol as for other models was followed here. The threshold was selected using the validation datasets and then locked in for the testing phase [23]. This indicates poor calibration. Under domain shifts and different datasets, this led to high recall on YawDD and DROZY with 0 specificity, showing that the low threshold is due to probability miscalibration rather than poor ranking. This is supported with the strong PR-AUC value [24]. The results for the final test show that the PR-AUC for CNN-GRU was 0.714 and 0.695 for UTA and DROZY respectively. This was the highest amongst all the models and this shows that it still has a strong ranking/capability despite the threshold miscalibration.Future work can implement post-hoc calibration methods such as temperature scaling on validation data to correct this without damaging the evaluation protocol [23].

D. Implementation in the Real-World

For ADAS deployment, based on the UTA F1-score, the R(2+1)D-18 model is the strongest on the internal benchmark (UTA-RLDD) with CNN-GRU and TSM closely trailing. In terms of the external testing and robustness CNN-GRU has better performance than both other models (Table VI); the model would just require some calibration before deployment. This is apt as well as the model requires a lower computational cost as it avoids full 3D convolutions unlike 3D CNN. The lightweight model has good performance and strong PR-AUC but is just not as good as the others. The current F1-scores (0.622–0.651) fall below the required thresholds for ADAS deployment and further optimisation is crucial for real-world implementation.

E. Limitations

Clip level detection of drowsiness is efficient for storing the clips and is currently better for a uniform evaluation process. This process does however cause only very short time periods to be analysed and drowsiness is a gradual process; it cannot just be quantified into a clip [26]. As thresholds were tuned using validation and froze them this meant that the CNN-GRU value was mis-calibrated. This could have been fixed by manual rectification, but it would damage the sanctity of the study. This can be fixed further. Lastly, yawning does not confirm that the person is drowsy. This could lead to a lot of false positives on a technicality that the person was yawning but is not drowsy. So the performance of the models on

YawDD was to test if it could detect yawning not drowsiness directly.

IX. Conclusion

The study achieved the objectives by comparing the baseline models and the deep-learning models under a uniform evaluation protocol. All experiments ensured subject-independent splitting and that the dataset clips were processed identically for all the models. The thresholds were selected only on validation data and then internal and external testing were conducted to test the capability of the model.

The evaluation was thorough as there were three datasets. UTA-RLDD was the primary internal benchmark, YawDD was a driving environment yawning dataset made to test the domain shift of a driver and DROZY was an external test made to test low-light drowsiness detection capabilities. The deep models consistently outperformed the baseline models under subject-independent and cross-domain evaluation. The 3D CNN had the strongest internal performance whereas the CNN-GRU was the strongest in terms of external robustness (DROZY). Currently, based on the results, CNN-GRU is the model which would be most apt to be implemented in the real-world due to its performance and low computational cost.

The results show that temporal modelling is capable of cross-domain generalization whereas threshold calibration is difficult under the varying environments and features. The set thresholds reflected on the accuracy of the calibration process. Future works could include data leakage safe calibration methods, optimised domain adaptation and longer temporal analysis periods to improve the robustness and efficiency of the applied model.

References

[1] AAA Foundation for Traffic Safety, "Drowsy Driving in Fatal Crashes, United States, 2017–2021," 2024.

[2] ROSPA, "Driver fatigue and road collisions," 2024.

[3] A. Kelkar *et al.*, "Revolutionizing the driving experience: Enhancing safety and comfort using ADAS," in *Proc. Int. Conf. Computing Communication Control and Automation*, 2023, pp. 1–4.

[4] D. Liu *et al.*, "Design of a driver fatigue detection and graded warning system based on multi-feature fusion," in *Proc. 2nd Int. Conf. Intelligent Transportation and Smart Cities*, vol. 13682, 2025, pp. 120–126.

[5] R. Ghoddoosian, M. Galik, and A. Vit, "A realistic dataset and baseline temporal model for early drowsiness detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 178–182.

[6] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, "YawDD: Yawning detection dataset," *IEEE Dataport*, 2020.

[7] Q. Massoz, T. Langohr, C. François, and J. G. Verly, "The ULg multimodality drowsiness database (called DROZY) and examples of use," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2016, pp. 1–7.

[8] Z. L. J. Huang, "Multi-feature fatigue driving detection based on computer vision," 2020.

[9] R. Kaur, Gitanjali, P. K. Awasthi, H. Dewangan, and S. Mehta, "DeepStress: CNN-GRU-based workplace stress detection using facial expressions and body posture," in *Proc. Int. Conf. Networks and Cryptology (NETCRYPT)*, 2025, pp. 811–815.

[10] J. Haddad, O. Lezoray, and P. Hamel, "3D-CNN for facial emotion recognition in videos," in *Proc. Int. Symp. Visual Computing (ISVC)*, 2020, pp. 298–309.

[11] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 7083–7093.

[12] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, p. 6086, 2024.

[13] T. Sundelin *et al.*, "Cues of fatigue: Effects of sleep deprivation on facial appearance," *Sleep*, vol. 36, no. 9, pp. 1355–1360, 2013.

[14] S. Essahraui *et al.*, "Real-time driver drowsiness detection using facial analysis and machine learning techniques," *Sensors*, vol. 25, no. 3, p. 812, 2025.

[15] D. Bhatt *et al.*, "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics*, vol. 10, no. 20, p. 2470, 2021.

[16] T. Schlosser, M. Friedrich, T. Meyer, and D. Kowerko, "A consolidated overview of evaluation and performance metrics for machine learning and computer vision," 2024.

[17] A. Georghiou and D. J. Bertsimas, "Binary decision rules for multistage adaptive mixed-integer optimization," *Springer*, 2017.

[18] A. Miley and Å. Torsvall, "Comparing two versions of the Karolinska Sleepiness Scale (KSS)," *Accident Analysis & Prevention*, vol. 94, pp. 134–141, 2016.

[19] O. Köpüklü, N. Köse, A. Gunduz, and G. Rigoll, "Resource efficient 3D convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Computer Vision Workshops (ICCVW)*, 2019, pp. 1910–1919.

[20] N. Kumar Rao B *et al.*, "Facial landmarks detection system with OpenCV Mediapipe and Python using optical flow (active) approach," in *Proc. 3rd Int. Conf. Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2023, pp. 92–96.

[21] D. Tran *et al.*, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6450–6459.

[22] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, pp. 1–35, 2016.

[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2017, pp. 1321–1330.

[24] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1521–1528.

[25] S. Desai and H. G. Ramaswamy, "Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2020, pp. 972–980.

[26] H. Mora, T. Echaveguren, and E. Pino, "Advanced forecasting of driver drowsiness events: Non-intrusive data and multimodal BiLSTM-based modeling," *Biomedical Signal Processing and Control*, 2026, Art. no. 109793.