SnoozeNet: An Ensemble CNN-MediaPipe Feature-based Pipeline with Temporal Convolutional Networks for Real-Time Driver Drowsiness Detection

Louie I. Calma Jr.
College of Computer Studies
Angeles University Foundation
Angeles City, Pampanga, Philippines

Christian Harvey G. Cayanan
College of Computer Studies
Angeles University Foundation
Angeles City, Pampanga, Philippines

Ian Carlo A. Reyes
College of Computer Studies
Angeles University Foundation
Angeles City, Pampanga, Philippines

Melissa Pantig
College of Computer Studies
Angeles University Foundation
Angeles City, Pampanga, Philippines

ABSTRACT
Driver drowsiness is a significant contributor to road accidents, often leading to impaired focus, delayed reaction times, and poor decision-making. To address this issue, this study introduces SnoozeNet, a lightweight and efficient real-time driver drowsiness detection system that combines Convolutional Neural Networks (CNNs), MediaPipe facial landmark tracking, and Temporal Convolutional Networks (TCNs). The model extracts spatial features from eye and mouth regions to detect blink rate, eye closure, and yawning, while MediaPipe provides head pose estimations to assess posture and nodding behavior. These features are fused and processed by a TCN to model behavioral transitions over time. The system was trained on diverse public datasets and evaluated against LSTM-based baselines, showing improved accuracy, training efficiency, and responsiveness. Results confirm that the lightweight CNN-MediaPipe-TCN pipeline effectively detects drowsiness-related facial cues across varied lighting conditions and facial structures, offering a robust and deployable solution for real-world driver-monitoring applications. Comprehensive validation showed that the pipeline achieved strong performance with an overall accuracy of 94.6%, F1-score of 0.930, and AUROC of 0.984, while delivering real-time classification in a browser-based application at approximately 15 FPS.

General Terms
Computer Vision, Deep Learning, Pattern Recognition, Real-time Systems

Keywords
Drowsiness detection, Temporal Convolutional Networks, MediaPipe, Convolutional Neural Networks, Real-time monitoring, Facial landmark detection, Deep learning, Computer vision.

1. INTRODUCTION
Driver drowsiness remains a major contributor to road accidents worldwide, impairing focus, reaction time, and decision-making in ways comparable to alcohol impairment [1, 2]. According to the World Health Organization, drowsiness-related crashes account for approximately 20% of all road traffic accidents globally [3]. In the Philippines, driver fatigue is recognized as a significant yet underreported factor in vehicular accidents [4]. Studies indicate that 67.3% of drivers have reported experiencing drowsiness while driving [3], with many commercial vehicle drivers reporting substantial fatigue during long-haul trips [2].

Recent advances in deep learning and computer vision have enabled the real-time monitoring of facial cues, such as blinking, yawning, and head movement, for drowsiness detection [7, 8]. However, most existing drowsiness detection models demand high computational power and expensive hardware, which limits their accessibility and practical deployment, especially in resource-constrained environments. Traditional approaches relied on intrusive physiological sensors or vehicle-based monitoring systems that suffered from high false-positive rates and delayed detection [2].

This study developed a lightweight ensemble Convolutional Neural Network-MediaPipe feature-based Temporal Convolutional Network (TCN) model to classify drowsy and non-drowsy states. It utilizes ensemble CNNs for spatial extraction of mouth and eye features, combined with MediaPipe facial landmark features and a TCN for real-time analysis of temporal behavior. By reducing hardware requirements, this new pipeline opens up opportunities for researchers in developing countries to develop and adapt drowsiness detection systems tailored to local needs.

2. RELATED WORK
2.1 Traditional Drowsiness Detection Methods
Early drowsiness detection systems relied on intrusive physiological sensors such as electroencephalography (EEG), electrocardiography (ECG), and electromyography (EMG) to monitor brain activity, heart rate, and muscle tension [2]. While these methods provided accurate measurements of fatigue-related physiological changes, they required direct contact with the driver, making them impractical for real-world deployment [5]. Vehicle-based detection methods emerged as non-intrusive alternatives, monitoring steering wheel movements, lane

departures, and vehicle speed variations as indicators of driver fatigue [2]. However, these approaches suffered from high false-positive rates, as driving behavior can be influenced by road conditions, traffic patterns, and driver experience rather than drowsiness alone.

2.2 Vision-Based Approaches

Recent advances in computer vision and deep learning have enabled non-intrusive, camera-based drowsiness detection systems. These systems analyze facial features such as eye closure patterns, blink frequency, yawning, and head pose to identify signs of fatigue [13, 14]. Convolutional Neural Networks (CNNs) have become the dominant approach for extracting spatial features from facial images, demonstrating superior performance in recognizing drowsiness-related visual patterns [15, 16]. MediaPipe, developed by Google, provides a robust framework for real-time facial landmark detection and head pose estimation [6]. Its lightweight architecture makes it suitable for deployment in resource-constrained environments, enabling efficient extraction of geometric features such as eye aspect ratio (EAR), mouth aspect ratio (MAR), and head orientation angles.

2.3 Temporal Modeling in Drowsiness Detection

Drowsiness manifests as temporal patterns rather than isolated events, requiring models capable of capturing sequential dependencies. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been widely used for temporal modeling in drowsiness detection [18, 19]. However, LSTMs suffer from vanishing gradient problems during training and limited parallelization capabilities, making them less efficient for real-time applications [7]. Temporal Convolutional Networks (TCNs) have emerged as a powerful alternative, offering parallelizable training, stable gradients, and flexible receptive fields through dilated convolutions [8]. TCNs have demonstrated superior performance in sequence modeling tasks, outperforming LSTMs in accuracy and training efficiency [9].

3. METHODOLOGY

3.1 System Architecture Overview

The proposed SnoozeNet system comprises three main components: CNN-based feature extraction for eye and mouth regions, MediaPipe-based head pose estimation, and a Temporal Convolutional Network for sequence modeling. The architecture is designed to balance computational efficiency with detection accuracy, enabling real-time deployment on standard hardware. The pipeline processes video frames at 15 FPS, extracting spatial and temporal features that are combined for drowsiness classification.

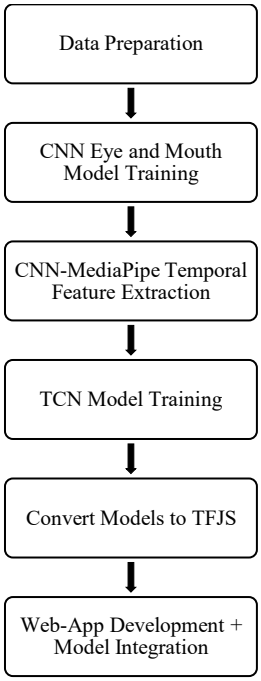

Figure 1. SnoozeNet Pipeline Architecture

3.2 Feature Extraction and Preprocessing

3.2.1 Facial Landmark Detection

MediaPipe FaceMesh was employed to extract 468 three-dimensional facial landmarks from each video frame. This provides robust spatial reference points for calculating eye, mouth, and head-pose features. Key landmark regions include eye landmarks (33-133 for left eye, 362-263 for right eye), mouth landmarks (61, 291, 13, 14, 81, 178, 308, 402), and head-pose anchors (outer eye corners at 33 and 263, chin at 152, forehead at 10).

3.2.2 Eye and Mouth Aspect Ratios

Eye Aspect Ratio (EAR) quantifies the degree of eye openness using six key landmarks around each eye. The formula computes the ratio of vertical eye-opening distances to horizontal eye span. EAR values decrease significantly when eyes close, providing a reliable indicator of eye state. Mouth Aspect Ratio (MAR) measures the relative vertical opening of the mouth and serves as a key indicator of yawning behavior. The formula computes the ratio of vertical mouth opening distances to horizontal mouth span. Higher MAR values indicate mouth opening, with sustained high values signaling yawning events.

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

$$MAR = \frac{||m_3 - m_9|| + ||m_4 - m_8|| + ||m_5 - m_7||}{3||m_1 - m_6||}$$

Figure 2. Eye Aspect Ratio and Mouth Aspect Ratio Illustration

3.2.3 Head Pose Estimation

Head pose estimation identifies the driver's yaw, pitch, and roll angles, which are critical for recognizing nodding or looking-away behaviors. The head orientation axes are derived from selected facial landmarks. A dynamic baseline calibration mechanism was implemented to account for varying camera positions and individual driver postures. The calibration process captures baseline head pose during the initial non-

drowsy state, allowing the system to detect deviations indicative of nodding behavior. Calibrated angles are computed by comparing current head orientation with the baseline rotation matrix, neutralizing camera-angle bias and subject-specific posture offsets.

3.2.4 Behavioral Feature Computation

Several temporal behavioral metrics were computed from the extracted facial features:

Eye Openness: CNN-predicted probabilities from the left and right eyes are combined. If the combined probability falls below 0.40, the eye is considered closed.

PERCLOS (Percentage of Eye Closure): Computed as the proportion of closed-eye frames within a 30-second sliding window, providing a standard fatigue metric.

Blink Detection: Characterized by short-duration eye closures lasting 2–6 frames (approximately 0.13–0.40 seconds). Blink rate over 30 seconds is tracked.

Eye Openness Trend: Short-term and long-term exponential moving averages (EMAs) are computed to capture gradual eye-closing behavior. The difference between them reflects the trend.

Yawning Detection: Mouth openness is derived from CNN-based yawn probability, smoothed using a 1-second EMA. A yawn event is triggered if the mouth remains open for at least 1.3 seconds.

Nod Detection: A nod event occurs when the head tilts downward (pitch exceeds threshold) while the eyes are closed, indicating fatigue-related head drops.

Table 1. Drowsy Event Formulas and Calculation Thresholds

Paramater	Condition / Value	Notes
Eye closed	$(p_{eye} < 0.40)$	–
Blink Duration	2-6 frames (0.13–0.40 s)	–
Prolonged closure	$(\geq 0.8 \times FPS) \approx 2.5s$	–
PERCLOS window	30 s (450 frames)	–
Mouth open	$(y_t \geq 0.55), close \leq 0.45$	Hysteresis
Yawn event	$(d_m \geq 1.3s)$	–
Head nod	$pitch \leq -4°, roll \leq 20°$	–
Sliding clip	6 s window, 3 s stride	–

3.3 CNN Architecture and Training

3.3.1 Eye Open/Closed CNN

A lightweight CNN was designed to classify eye state from grayscale 90×90 pixel cropped eye regions. The architecture consists of three convolutional blocks with progressively increasing filter sizes (32, 64, 128). Each convolutional layer uses 3×3 kernels followed by batch normalization, ReLU activation, and 2×2 max pooling. The flattened features are processed through dense layers with dropout regularization (rate=0.3 - 0.2) before the final sigmoid output layer for binary classification.

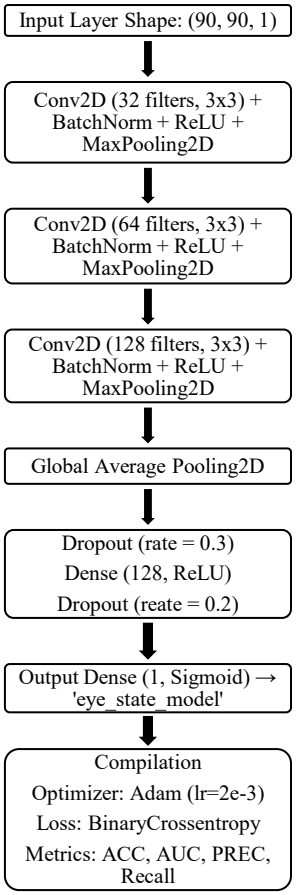

Figure 3. CNN Architecture Parameters - Eye Open/Closed Model

Training was performed using the Adam optimizer with binary cross-entropy loss. The learning rate was set to 0.001 with a decay schedule. Data augmentation techniques including random rotation (±15°), brightness adjustment (±20%), and horizontal flipping were applied to improve generalization. The model was trained for 30 epochs with early stopping (patience=10) on validation loss. The dataset split was 70% training, 15% validation, and 15% testing, ensuring sufficient data for learning while maintaining unbiased evaluation.

3.3.2 Mouth Open/Closed CNN

To accommodate the finer spatial details of the mouth region, the mouth CNN processes larger 120×120 grayscale inputs. The architecture employs four convolutional blocks with filter sizes (32, 64, 128, 256) to capture complex mouth opening patterns associated with yawning. Each block includes batch normalization and dropout (rate=0.5) for regularization. The network architecture is deeper than the eye CNN to better capture the wider range of mouth shapes and yawning expressions.

The mouth CNN was trained using Adam optimizer with binary cross-entropy loss over 40 epochs. The same data augmentation strategy was applied as for the eye CNN. Training utilized frames extracted from YawDD and D3S datasets, with MAR-based auto-labeling followed by manual validation to ensure label quality. The model successfully learned to distinguish between closed mouth, open mouth, and yawning states with high confidence.

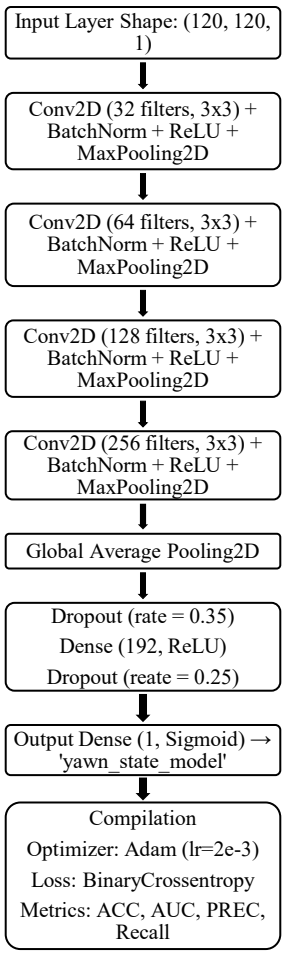

Figure 4. CNN Architecture Parameters - Mouth Open/Closed Model

3.4 Temporal Convolutional Network Architecture

The Temporal Convolutional Network (TCN) was designed to capture long-range temporal dependencies in drowsiness behavior patterns. The architecture processes sequences of 90 frames (6 seconds at 15 FPS) using dilated 1D causal convolutions. The dilation rates follow an exponential pattern (1, 2, 4, 8), enabling the network to capture dependencies across different temporal scales without significantly increasing model parameters.

Each TCN block consists of two dilated convolutional layers with 64 filters and kernel size 3, followed by batch normalization, ReLU activation, and spatial dropout (rate=0.3). Residual connections are employed to facilitate gradient flow and enable deeper architectures. The TCN receives concatenated features including CNN probability outputs (eye openness, yawn probability), MediaPipe-derived pose angles (calibrated yaw, pitch, roll), and computed behavioral metrics (PERCLOS, blink rate, eye closure duration).

Training employed a focal loss to address class imbalance between drowsy and non-drowsy frames, with alpha=0.75 and gamma=2.0. The Adam optimizer was used with an initial learning rate of 0.0001 and cosine annealing scheduling. Early stopping was applied with patience of 15 epochs monitoring validation AUROC. A sliding window approach with 3-second stride was used to create temporal sequences, with coarse labels (binary drowsy/non-drowsy) assigned based on the presence of any drowsiness indicator within the window.

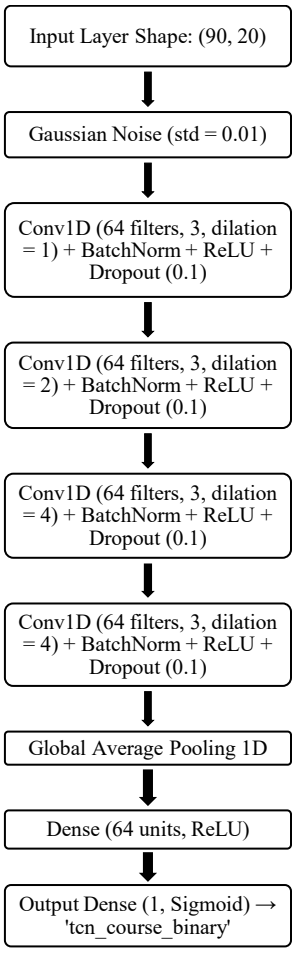

Figure 5. TCN Architecture Parameters and Hyperparameters

3.5 LSTM Baseline Architecture

For comparative evaluation, an LSTM-based architecture was implemented as a baseline. The LSTM processes the same 90-frame input sequences with 64 hidden units per layer. The architecture consists of two stacked LSTM layers with dropout (rate=0.25) for regularization. The final LSTM output is processed through dense layers with ReLU activation before the sigmoid output layer. Training utilized the same optimizer, loss function, and learning rate scheduling as the TCN for fair comparison.

3.6 Datasets

The system was trained and evaluated on multiple public datasets to ensure generalization across diverse demographics and conditions: NTHU Drowsy Driver Detection (NTHU-DDD), which contains infrared videos with detailed annotations on eye states and yawning [10], YawDD focused on yawning detection with diverse lighting and head positions [11], Driver Drowsiness Dataset (D3S) providing annotated video sequences with multiple drowsiness indicators [12], the DMD Dataset, offering large-scale multi-modal driver monitoring data [13], and Open-Closed Eyes Dataset specialized for eye state classification [14].

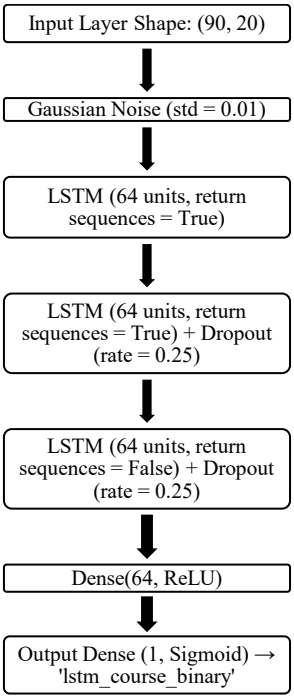

Figure 6. LSTM Architecture Parameters and Hyperparameters

3.7 Evaluation Metrics and Validation

Model performance was evaluated using standard classification metrics: accuracy measuring overall correctness, precision indicating the proportion of true positives among predicted positives, recall measuring the proportion of actual positives correctly identified, F1-score representing the harmonic mean of precision and recall, AUROC (Area Under Receiver Operating Characteristic curve) assessing discrimination ability across all thresholds, and AUPRC (Area Under Precision-Recall Curve) particularly important for imbalanced datasets.

Two validation strategies were employed: holdout validation using an 80-20 train-test split for overall performance assessment, and Leave-One-Subject-Out (LOSO) cross-validation where each fold held out all data from one driver for testing while training on the rest. LOSO validation effectively prevents identity leakage and enables realistic deployment assessment, evaluating the model's ability to generalize to completely unseen individuals without requiring personalized calibration.

4. RESULTS AND DISCUSSION

4.1 CNN Feature Extraction Performance

The CNN models successfully generated stable and discriminative probability outputs for eye and mouth regions. The eye CNN achieved 97.8% accuracy in distinguishing between open and closed eyes, with strong generalization across different lighting conditions and facial orientations. The mouth CNN demonstrated 96.4% accuracy in detecting yawning, effectively handling variations in mouth shape and head pose. These probability outputs remained stable and responsive to facial state transitions while maintaining low noise during alert phases.

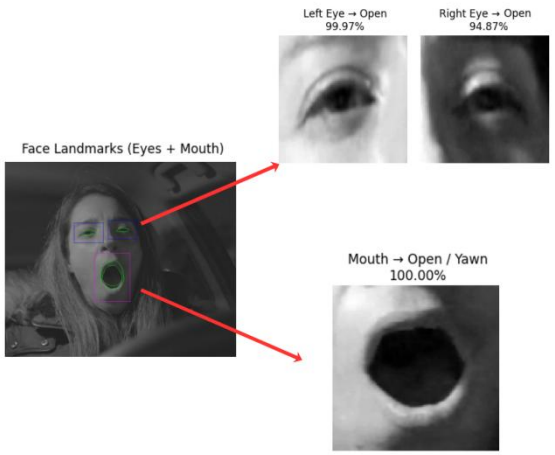

Figure 7. CNN Probability Output Visualization

The yaw-pitch-roll calibration mechanism significantly improved the reliability of pose features by neutralizing camera-angle bias and subject-specific posture offsets. As demonstrated in evaluation, the calibrated orientation signals remained centered and consistent across sessions, allowing the TCN to interpret head movements as genuine behavioral patterns rather than environmental artifacts. This calibration was particularly effective in distinguishing between natural head movements during normal driving and nodding behavior indicative of drowsiness.

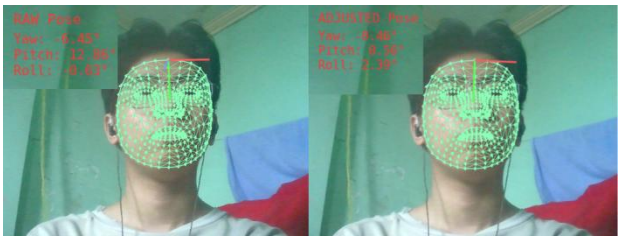

Figure 8. Calibrated Head Pose Angles (Yaw, Pitch, Roll)

4.2 Training Efficiency Analysis

The proposed feature-based pipeline demonstrated high training efficiency across different hardware configurations. CNN training for eye detection required 40 minutes on GTX 1050 and 9 minutes on RTX 2060 Super, while mouth detection training took 34 minutes and 14 minutes respectively. These results demonstrate that the CNN-based approach can be executed on lower-end hardware, though with considerably slower training compared to higher-tier GPUs.

Despite the computational demands of sequence modeling, both TCN and LSTM frameworks trained rapidly due to the efficiency of the feature-based pipeline. TCN training required only 6 seconds on GTX 1050 and 9 seconds on RTX 2060 Super, while LSTM training took 15 seconds and 6 seconds, respectively. The TCN demonstrated notably faster training times, indicating that the optimized architecture minimizes processing overhead and enables efficient training of resource-intensive temporal models across different hardware configurations.

Table 2. CNN Training Time Comparison Across Hardware

GPU	Yawn Model	Blink Model
GTX 1050	34m 15s	40m 46s
RTX 2060 Super	14m 30s	9m 17s

Table 3. Temporal Model Training Time Comparison

GPU	Yawn Model	Blink Model
GTX 1050	9s	15s
RTX 2060 Super	6s	9m 17s

4.3 Overall Performance Comparison

The CNN-MediaPipe-TCN pipeline achieved superior performance compared to the LSTM baseline across all evaluation metrics. The TCN model achieved an overall accuracy of 94.6%, F1-score of 0.930, AUROC of 0.984, and AUPRC of 0.980. The model demonstrated strong precision (0.91) and recall (0.95) for the drowsy class, indicating effective detection of fatigue states with low false-positive rates. For the non-drowsy class, the model achieved precision of 0.97 and recall of 0.95, confirming reliable classification of alert states.

Table 4. Per-Class Evaluation Metrics for CNN-MediaPipe-TCN Model

Class	Precision	Recall	F1-Score	Support
Awake	0.97	0.95	0.96	91
Drowsy	0.91	0.95	0.93	56

In contrast, the LSTM-based model achieved lower performance, with an accuracy of 89.8%, F1-score of 0.870, AUROC of 0.950, and AUPRC of 0.931. While LSTM was able to model temporal sequences, it showed reduced precision in detecting drowsy states (0.85 vs 0.91 for TCN), suggesting a higher rate of false positives. The recall for drowsy detection was also lower (0.89 vs 0.95), indicating missed drowsiness events. These results demonstrate the TCN's superior ability to capture temporal dependencies in drowsiness behavior while maintaining higher classification accuracy.

Table 5. Per-Class Evaluation Metrics for CNN-MediaPipe-LSTM Model

Class	Precision	Recall	F1-Score	Support
Awake	0.93	0.90	0.92	91
Drowsy	0.85	0.89	0.87	56

A direct comparison of overall performance metrics confirms the TCN's advantages. The TCN model outperformed LSTM in accuracy (94.6% vs 89.8%), F1-score (0.930 vs 0.870), and AUROC (0.984 vs 0.950). The performance gap is particularly evident in the AUROC metric, with TCN showing 3.4 percentage points higher discrimination ability. This superior performance can be attributed to TCN's architectural advantages: parallel training enabling stable gradients, dilated convolutions providing exponentially growing receptive fields for capturing long-range dependencies, and reduced memory burden compared to maintaining LSTM hidden states.

Table 6. Overall Performance Comparison: TCN vs LSTM

Metric	TCN	LSTM
Accuracy	0.946	0.898
AUROC	0.984	0.950
F1	0.93	0.87
Total samples	147	147

4.4 Subject-Independent Generalization

Leave-One-Subject-Out (LOSO) cross-validation was performed to evaluate subject-independent generalization, a critical requirement for real-world deployment. The CNN-MediaPipe-TCN pipeline demonstrated strong generalization with average metrics of ACC = 0.882, F1 = 0.837, AUROC = 0.939, and AUPRC = 0.938 across all subjects. These results show robust precision-recall behavior across varying class prevalence and confirm the model's ability to generalize to unseen drivers without requiring individual calibration.

The LOSO validation revealed consistent performance across most subjects, with some variation depending on individual drowsiness manifestation patterns. Subjects with more pronounced drowsiness indicators (e.g., longer eye closure durations, frequent yawning) showed higher detection accuracy, while those with subtle or atypical patterns presented greater challenges. Despite this variability, the average performance remained high, demonstrating the model's robustness to inter-subject differences.

Table 7. LOSO Cross-Validation Results - CNN-MediaPipe-TCN Model

Subject	ACC	F1	AUROC	AUPRC	Pos(%)
person001	0.980	0.985	1.000	1.000	64.71
person002	0.837	0.818	0.956	0.954	44.90
person003	0.843	0.789	0.981	0.982	45.10
person004	0.840	0.840	0.915	0.939	48.00
person005	0.936	0.941	0.985	0.990	55.32
person006	0.960	0.964	0.994	0.996	56.44
person007	0.833	0.818	0.911	0.936	45.83
person008	0.961	0.875	0.994	0.975	15.69
person009	0.882	0.870	0.992	0.988	39.22
person010	0.837	0.882	0.921	0.980	77.55
person011	0.796	0.722	0.907	0.912	44.90
person012	0.981	0.984	0.997	0.998	58.49
person013	0.809	0.471	0.652	0.574	25.53
person014	0.857	0.759	0.945	0.918	32.65
Average	**0.882**	**0.837**	**0.939**	**0.938**	**46.74**

In contrast, the CNN-MediaPipe-LSTM baseline achieved lower LOSO performance with averages of ACC 0.827, F1 0.774, AUROC 0.897, and AUPRC 0.884. The LSTM model showed greater sensitivity to subject-specific variability, particularly for profiles with low positive rates or atypical drowsiness patterns. Several subjects showed significantly degraded performance with LSTM, indicating reduced generalization capability compared to TCN. The F1-score difference of 0.063 (0.837 vs 0.774) is particularly notable, confirming TCN's superior balance between precision and recall across diverse individuals.

Table 8. LOSO Cross-Validation Results - CNN-MediaPipe-LSTM Model

Subject	ACC	F1	AUROC	AUPRC	Pos(%)
person001	0.667	0.653	0.847	0.927	64.71
person002	0.837	0.810	0.949	0.944	44.90
person003	0.941	0.933	0.960	0.949	45.10
person004	0.680	0.680	0.737	0.734	48.00
person005	0.872	0.897	0.963	0.971	55.32
person006	0.921	0.926	0.972	0.983	56.44
person007	0.854	0.829	0.904	0.921	45.83
person008	0.902	0.545	0.913	0.738	15.69
person009	0.725	0.741	0.950	0.923	39.22
person010	0.755	0.829	0.861	0.957	77.55
person011	0.776	0.703	0.843	0.866	44.90
person012	0.906	0.921	0.977	0.984	58.49
person013	0.809	0.471	0.748	0.555	25.53
person014	0.939	0.897	0.936	0.929	32.65
Average	**0.827**	**0.774**	**0.897**	**0.884**	**46.74**

These LOSO results align with expectations that temporal convolutions are more resilient to long-range variance and training instability than recurrent models, especially under limited data conditions. The TCN's ability to maintain consistent performance across all subjects without personalized calibration demonstrates its practical viability for commercial deployment in driver monitoring systems.

4.5 Real-Time Deployment Performance

The optimized lightweight pipeline was successfully deployed as a browser-based web application using TensorFlow.js, enabling real-time drowsiness detection via a standard webcam without specialized hardware. The system was evaluated on consumer-grade hardware (Intel Core i5, 8GB RAM, integrated graphics) to assess practical deployment viability. The pipeline achieved real-time inference at approximately 15 FPS with average latency of 67ms per frame, demonstrating sufficient responsiveness for driver monitoring applications.

The web application implements a severity classification system based on accumulated drowsiness duration thresholds. The system classifies the driver as non-drowsy when total drowsy time is below 1 second, drowsy when exceeding 1 second, and critical when surpassing 5 seconds. This threshold-based severity assessment provides graduated warnings, allowing drivers to recognize fatigue progression before reaching critical levels. The system successfully detected prolonged eye closure, yawning events, and nodding behavior in real-time testing scenarios across different users and lighting conditions.

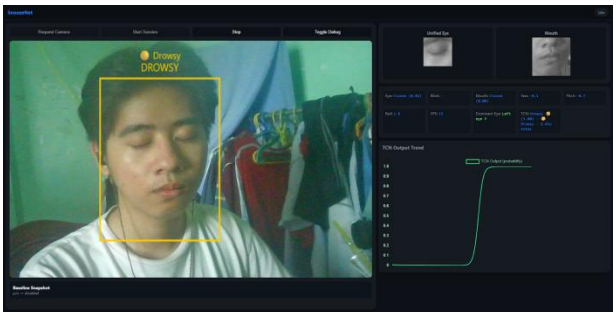

Figure 9. SnoozeNet Web Application Interface - Drowsiness Detection

4.6 Discussion

The superior performance of the TCN-based approach can be attributed to several architectural advantages over LSTM. Unlike LSTMs which process sequences recurrently and suffer from vanishing gradients, TCNs employ causal dilated convolutions that enable parallel training and stable gradient flow. The dilated convolutions provide exponentially growing receptive fields, allowing the model to capture long-range temporal dependencies without the memory burden of maintaining hidden states across all time steps.

The integration of MediaPipe head pose estimation proved particularly valuable for detecting nodding behavior, a subtle but reliable indicator of drowsiness. The calibration mechanism successfully accounted for inter-individual differences in neutral head position, reducing false positives caused by natural head movements during normal driving. The combination with CNN-based eye and mouth analysis ensured robust detection even when only one indicator (e.g., eye closure without nodding, or yawning without sustained eye closure) was present.

The feature-based pipeline's computational efficiency enables deployment on standard consumer hardware without GPU acceleration. By extracting compact behavioral features (probabilities, angles, rates) rather than processing raw video frames through the temporal model, the system significantly reduces memory requirements and enables real-time inference in browser environments. This accessibility is crucial for widespread adoption, particularly in developing countries where specialized hardware may not be readily available.

The strong LOSO cross-validation results demonstrate practical deployment viability. Unlike systems requiring personalized calibration or training data for each driver, SnoozeNet can be immediately deployed for new users with consistent performance. This characteristic addresses a critical barrier to commercial adoption, as fleet operators can implement the system across diverse driver populations without individual setup procedures.

5. CONCLUSION

This study successfully developed SnoozeNet, an ensemble CNN-MediaPipe-TCN pipeline for real-time driver drowsiness detection. The system achieves 94.6% accuracy with strong generalization to unseen drivers (LOSO ACC 0.882), outperforming LSTM-based approaches in both accuracy and computational efficiency. The integration of CNN-based feature extraction, MediaPipe head pose estimation, and Temporal Convolutional Networks provided a comprehensive solution that captures both spatial and temporal aspects of drowsiness behavior.

The lightweight architecture enables real-time deployment on standard consumer hardware at 15 FPS, making advanced drowsiness detection accessible for resource-constrained environments. The browser-based implementation using TensorFlow.js demonstrates practical viability for commercial deployment without specialized hardware requirements. The calibration mechanism for head pose and the ensemble approach for facial feature detection ensure robustness across varying camera angles, lighting conditions, and facial structures.

Future work will focus on several areas, including: expanding the training dataset with more diverse demographics and environmental conditions, particularly nighttime and low-light scenarios; integrating multi-modal signals such as steering patterns, vehicle performance data, and physiological sensors for enhanced detection reliability; optimizing the model for edge deployment in embedded automotive systems through quantization and pruning techniques; and exploring attention mechanisms or transformer architectures for potentially improved temporal modeling. Additionally, long-term field studies with commercial drivers would provide valuable insights into real-world performance and user acceptance.

6. ACKNOWLEDGMENTS

The researchers acknowledge the support provided by Angeles University Foundation and the guidance of their thesis adviser throughout this research. They also thank the creators of the public datasets used in this study for making their data available to the research community.

7. REFERENCES

[1] Chen, L., Xin, G., Liu, Y., & Huang, J. 2021. Driver fatigue detection based on facial key points and LSTM. Security and Communication Networks, 2021(1), 5383573.

[2] Deshpande, A. V., Kalambarkar, S., Kale, A., Joshi, R., Kadam, S., Kadam, V., & Kale, N. 2024. A hybrid approach to drowsiness detection using MediaPipe and YOLOv5. International Journal of Innovative Research in Technology (IJIRT), 11(6), 3556–3561.

[3] Van Houdt, G., Mosavi, A., Napoles, G., & Shahriari-Rad, M. 2020. A review on the long short-term memory model. AI, 1(1), 110–122.

[4] Rundo, F., Spampinato, C., & Rundo, M. 2023. Car-driver drowsiness assessment through 1D temporal convolutional networks. arXiv preprint arXiv:2308.02415.

[5] Ismail, A. R., Sodoyer, D., & Elbahhar Boukour, F. 2023. Drowsiness detection in humans based on ECG analysis using temporal convolutional network. In Proceedings of the 2023 International Conference on Automation, Control, and Electrical Engineering (CACEE). IEEE.

[6] Weng, C. H., Lai, Y. H., & Lai, S. H. 2016. NTHU Drowsy Driver Detection Dataset.

[7] Abtahi, S., Omidyeganeh, M., Shirmohammadi, S., & Hariri, B. 2014. YawDD: A Yawning Detection Dataset. IEEE DataPort.

[8] Bindu, J., Jose, B. M., & Anjali, V. 2020. Driver Drowsiness Dataset (D3S). GitHub.

[9] Ortega, J. D., Kose, N., Cañas, P., Chao, M.-A., Unnervik, A., Nieto, M., Otaegui, O., Salgado, L. 2020. DMD: A Large-Scale Multi-modal Driver Monitoring Dataset for Attention and Alertness Analysis. In: A. Bartoli & A. Fusiello (eds), Computer Vision — ECCV 2020 Workshops. Springer International Publishing.

[10] Mammadli, S. 2024. Open and Closed Eyes Dataset.