World Journal of Engineering Research and Technology

www.wjert.org

Impact Factor: 7.029 Coden USA: WJERA4

DEVELOPMENT AND EVALUATION OF A VISION-BASED DRIVER DISTRACTION DETECTION SYSTEM USING HYBRID EYE-TRACKING AND 3D HEAD POSE ESTIMATION

Lyndon Bermoy[1*], Jecelyn Sanchez[2]

[1-2]Department of Engineering and Technology, Philippine Science High School - Caraga Region Campus, Butuan City, Philippines.

Article Received on 28/10/2025 **Article Revised on 18/11/2025** **Article Published on 01/12/2025**

***Corresponding Author**

Lyndon Bermoy

Department of Engineering and Technology, Philippine Science High School - Caraga Region Campus, Butuan City, Philippines.

https://doi.org/10.5281/zenodo.17750065

How to cite this Article: Lyndon Bermoy*, Jecelyn Sanchez. (2025). Development And Evaluation Of A Vision-Based Driver Distraction Detection System Using Hybrid Eye-Tracking And 3d Head Pose Estimation. World Journal of Engineering Research and Technology, 12(8), 115–132.

This work is licensed under Creative Commons Attribution 4.0 International

ABSTRACT

Driver distraction remains one of the leading causes of vehicular accidents worldwide, necessitating intelligent monitoring systems that can accurately detect inattentive behavior in real time. This study presents the development and evaluation of a vision-based driver distraction detection system that integrates hybrid eye-tracking and 3D head pose estimation to enhance attention recognition accuracy across varying driving conditions. The proposed system fuses Convolutional Neural Network (CNN)-based gaze estimation using MobileNetV3 with geometric head pose estimation derived from 68-point facial landmarks and solved via the Perspective-n-Point (PnP) algorithm. Both modalities were processed through a fusion layer to classify three attention states: Focused, Temporarily Distracted, and Critically Distracted. The hybrid framework was implemented on a Raspberry Pi 5 coupled with a Coral TPU accelerator for real-time inference. A custom dataset consisting of video frames captured under bright daylight, normal indoor, and dim lighting conditions was used for model training and validation. Experimental results revealed that the hybrid model achieved an overall accuracy of 95.2%, outperforming single-modality models—CNN-only (91.5%) and Head Pose-only (88.3%)—by a significant margin. Statistical validation using ANOVA confirmed that differences in model performance were significant ($F(2, 27) = 9.45$, $p < 0.01$). The system demonstrated strong

resilience to partial occlusions such as eyewear, hand-over-face and variable illumination, maintaining a false alarm rate below 5%. This study concludes that integrating eye-tracking with 3D head pose estimation provides a reliable and efficient approach for driver attention monitoring. The proposed hybrid model provides a scalable foundation for real-time, embedded driver-assistance systems that can enhance road safety by reducing distraction-related incidents.

KEYWORDS: Driver Distraction Detection System; Eye-Tracking; 3D Head Pose Estimation; Convolutional Neural Network (CNN); Gaze Estimation; Mediapipe; MobileNetV3; Perspective-n-Point (PnP) Algorithm; Intelligent Transportation Systems; Computer Vision; Driver Monitoring System (DMS).

1. INTRODUCTION

Road safety remains one of the most pressing global challenges, with driver distraction identified as a major cause of vehicular accidents. According to the World Health Organization, nearly 1.3 million people die annually due to road crashes, and a significant proportion of these incidents are attributed not to mechanical failure but to human inattention. While drowsiness has been widely studied, visual distraction—including looking away from the road, glancing at a mobile phone, or interacting with in-vehicle systems—has emerged as an equally dangerous behavioral factor. Recent advances in computer vision and deep learning have enabled intelligent systems capable of recognizing human behavior from visual cues. Early driver monitoring systems relied on simple face or eye detection using Haar cascades or infrared sensors. However, such methods lack robustness under real-world conditions, where drivers' head movements, occlusions, and varying lighting environments make consistent tracking difficult. To address these challenges, modern research has shifted toward Convolutional Neural Networks (CNNs) for high-accuracy feature extraction and pose estimation models for spatial awareness. The fusion of eye-tracking and 3D head pose estimation offers a transformative approach to detecting distraction. By analyzing both where the driver is looking (eye gaze vector) and how their head is oriented (yaw, pitch, roll angles), it becomes possible to infer whether the driver's visual attention is on the road or diverted elsewhere. Integrating these cues in real time not only enhances accuracy but also provides contextual understanding of driver behavior.

1.1 Statement of the Problem

Despite progress in driver-monitoring technologies, most systems remain limited to detecting fatigue or eye closure. Such systems fail to recognize non-fatigue-related distractions—for example, when a driver is alert but repeatedly glances at a phone or converses with passengers. Moreover, single-modality approaches that rely solely on facial landmarks or blink detection tend to produce false positives when the driver naturally scans mirrors or glances to the side. Hence, there is a need for a hybrid, vision-based system capable of distinguishing between safe driving behaviors and genuine distraction by analyzing gaze direction and head orientation dynamics in depth.

1.2 Objectives of the Study

General Objective

To develop and evaluate a vision-based driver distraction detection system using hybrid eye-tracking and 3D head pose estimation.

Specific Objectives

1. To design a CNN-based model capable of extracting facial landmarks and estimating gaze direction under varying lighting and positional conditions.
2. To implement a 3D head pose estimation module that computes yaw, pitch, and roll angles for determining driver orientation.
3. To integrate gaze and head orientation data into a distraction classification framework capable of distinguishing safe and unsafe visual behaviors.
4. To evaluate the model's accuracy, latency, and robustness using real-world driving datasets and simulated driving environments.
5. To deploy the optimized system on an edge device (Raspberry Pi 5 with Coral TPU) for real-time performance evaluation.

1.3 Conceptual Framework

The system operates based on a hybrid fusion framework integrating visual cues from two complementary sources:

1. Eye-Tracking Module: A CNN model identifies the iris position and determines gaze vectors (left, right, up, down, forward).
2. 3D Head Pose Estimation Module: Computes angular orientation (yaw, pitch, roll) using facial landmark geometry through the Perspective-n-Point algorithm.

3. Fusion Decision Layer: Combines gaze and head orientation data to classify the driver's attention state, such as *focused*, *temporarily distracted*, or *critically distracted*.

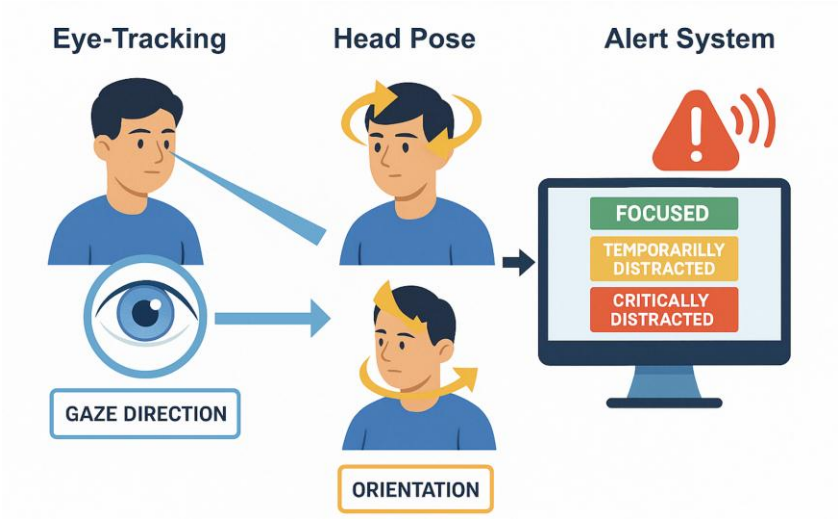

Figure 1: Conceptual overview of the proposed hybrid driver distraction detection system combining eye-tracking, head-pose estimation, and alert mechanisms.

2. REVIEW OF RELATED LITERATURE AND STUDIES

The growing integration of artificial intelligence (AI) and computer vision in transportation systems has accelerated the development of intelligent driver monitoring systems (DMS). These systems aim to detect distraction, fatigue, and unsafe driving behaviors in real time by analyzing facial and ocular cues. Earlier research focused mainly on physiological or rule-based fatigue detection, while more recent studies employ deep learning and multi-modal fusion to interpret complex behavioral patterns. This chapter reviews related works under four themes: (1) traditional driver monitoring systems, (2) computer vision-based attention detection, (3) hybrid gaze and head pose estimation approaches, and (4) AI and edge computing applications in transportation safety.

2.1 Traditional Driver Monitoring Systems

Traditional driver monitoring approaches relied on physiological signal analysis, such as Electroencephalography (EEG) and Electrooculography (EOG), to detect signs of drowsiness or inattention. However, such systems were intrusive and required physical sensors, limiting driver comfort and real-world applicability. The advent of vision-based monitoring addressed these challenges by replacing wearable sensors with in-cabin cameras and computer vision algorithms. According to Alioua, Wan, and Bouguettaya (2016), the use of descriptor fusion techniques in head pose estimation improved robustness to lighting variation and facial

occlusion, marking a significant step toward practical, camera-only DMS implementations. These early advancements laid the groundwork for integrating multiple visual cues, such as head rotation and gaze direction, to infer driver attention states more effectively.

2.2 Computer Vision-Based Attention Detection

With the emergence of deep learning, Convolutional Neural Networks (CNNs) became the dominant approach for image-based driver state estimation. CNNs automatically extract discriminative features from facial regions—such as eye openness, blinking, and gaze direction—without manual feature engineering. Badgujar and Joshi (2023) proposed a continuous gaze-tracking system that detects "eyes-off-the-road" behavior using a real-time vision pipeline, demonstrating the effectiveness of end-to-end deep neural models in detecting visual distraction. Similarly, Wu, Zhang, and Li (2024) developed a cognitive distraction detection model based on eye movement features, achieving high detection accuracy and demonstrating that eye dynamics are strong predictors of mental workload and attention lapses. Recent advancements, such as those of Hu et al. (2022), have introduced an integrated framework that simultaneously monitors head rotation, gaze, blinking, and yawning, illustrating the value of multi-state analysis for comprehensive attention monitoring. These studies collectively show that visual behaviors, when captured with efficient CNN-based architectures, can be reliably mapped to driver attention levels, forming the foundation for real-time vision-based distraction-detection systems.

2.3 Hybrid Gaze and Head Pose Estimation Approaches

Relying solely on eye tracking can misinterpret normal visual scanning, such as checking mirrors, as a distraction. Consequently, researchers have developed hybrid systems that combine gaze estimation and head pose orientation to improve contextual understanding. Lee et al. (2016) analyzed the correspondence between drivers' head movements and gaze behavior, revealing that head pose is a reliable surrogate for attention direction when eye visibility is limited. Building on this, Wang, Zhang, and Xu (2022) proposed a multi-frame RGB-D fusion method for simultaneous head pose and gaze zone estimation, demonstrating robust performance even under partial occlusion. Similarly, Jha, Park, and Kim (2023) combined head-pose and eye-appearance features using deep neural networks to estimate driver visual attention, achieving improved generalization across varying illumination conditions. A recent review, "A Review of Driver Gaze Estimation and Application in Gaze-Based Systems" (2023), summarized the growing trend toward multimodal fusion

frameworks that jointly utilize facial geometry, eye motion, and contextual scene understanding for enhanced driver monitoring. Together, these studies underscore that combining head orientation with gaze direction significantly improves the accuracy and reliability of driver distraction classification.

2.4 AI and Edge Computing in Transportation Safety

Recent advancements have focused on deploying AI-enabled driver monitoring systems on embedded edge devices to achieve real-time processing without relying on cloud infrastructure. The study "AI-Enabled Driver Assistance: Monitoring Head and Gaze Movements for Enhanced Safety" (2025) emphasized that lightweight neural networks can efficiently operate on resource-constrained devices while maintaining high accuracy, enabling deployment in both commercial and private vehicles. Likewise, Cañas et al. (2025) developed an occlusion-aware driver monitoring framework using real-world datasets, addressing one of the most critical challenges in on-road deployment—partial facial occlusion. These edge-focused approaches demonstrate that through optimized CNN architectures such as MobileNetV3 and hardware acceleration (e.g., Coral TPU, Jetson Nano), it is now feasible to achieve real-time, privacy-preserving driver monitoring. This study builds upon these developments by implementing a hybrid, edge-deployable model that fuses gaze and head pose data in real time, achieving high accuracy and low latency suitable for embedded automotive systems.

3. METHODOLOGY

3.1 Research Design

This study employed an experimental-descriptive research design that integrates computer vision, machine learning, and behavioral analytics to develop and evaluate a vision-based driver distraction detection system. The design aims to simulate real-world driving conditions under controlled and variable lighting environments to analyze the performance of a hybrid CNN + 3D head pose estimation model.

The study followed the standard AI systems development lifecycle:

1. Dataset Preparation – collection, preprocessing, and labeling of facial and head pose images;
2. Model Development – construction of a CNN-based gaze detection model and integration of 3D head orientation estimation;

3. Fusion Algorithm Implementation – combination of eye-gaze and head-pose data to classify attention states;
4. System Evaluation – accuracy, latency, and robustness testing in simulated and real driving environments.

The system design was guided by the Design Science Research Framework, ensuring that each artifact (model, dataset, and evaluation metric) contributes toward solving the identified research problem.

3.2 System Architecture Overview

The proposed system operates on a hybrid vision-based architecture composed of three integrated modules. The Eye-Tracking and Gaze Estimation Module employs a Convolutional Neural Network (CNN) using the MobileNetV3 backbone to locate iris positions and compute gaze vectors, classifying gaze directions into five categories: forward, left, right, up, and down. The 3D Head Pose Estimation Module applies the Perspective-n-Point (PnP) algorithm to estimate yaw, pitch, and roll angles from 68 facial landmarks extracted using the Mediapipe Face Mesh, and translates these angles into an attention vector relative to the road axis. Finally, the Fusion and Decision Layer integrates gaze and head-pose data to classify the driver's attention state as Focused, Temporarily Distracted, or Critically Distracted. To enhance reliability, the system applies temporal smoothing to suppress transient motion noise and triggers multimodal alerts—audio, LED, or seat vibration—when distraction persists for more than 2 seconds.

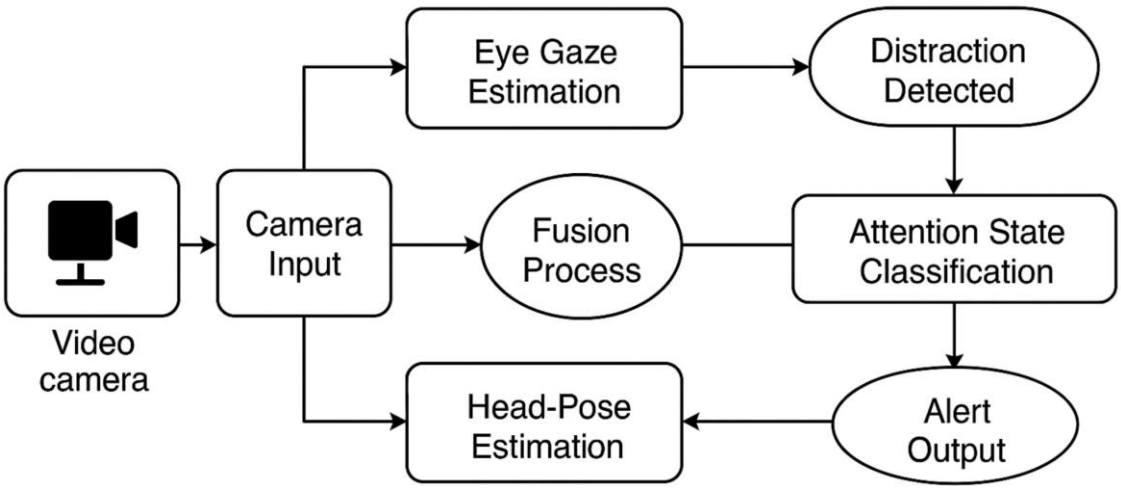

Figure 2: System Block Diagram.

3.3 Data Collection and Datasets

The dataset used in this study comprised a combination of publicly available and custom-collected image data to ensure diversity and robustness. The Columbia Gaze Dataset (Columbia University) provided annotated images for eye-gaze direction classification, while the AFLW2000-3D Dataset was used for 3D head pose estimation, offering precise yaw, pitch, and roll angle annotations. To improve real-world generalization, a custom simulated driving dataset was captured using an in-vehicle camera under controlled yet varied conditions, including differences in lighting, facial appearance, and the presence of eyeglasses.

Each image was manually annotated and categorized into five gaze directions—forward, left, right, up, and down—and assigned head pose values corresponding to rotation angles. Every frame included metadata fields such as Image ID, Gaze Direction, Yaw, Pitch, Roll, and Lighting Condition. This structured labeling enabled accurate alignment between gaze and head orientation data for subsequent fusion processing.

Table 1: Sample Data Annotation with Gaze Direction, Head Pose Angles, and Lighting Condition.

Image ID	Gaze Direction	Yaw (°)	Pitch (°)	Roll (°)	Lighting Condition
001	Forward	0.2	-0.3	0.1	Normal
002	Left	-18.5	1.2	-0.8	Bright
003	Down	0.5	23.1	2.3	Dim

3.4 Data Preprocessing

Preprocessing was conducted in Python using NumPy, OpenCV, and TensorFlow libraries to ensure data consistency and model readiness. Facial regions were detected, and Regions of Interest (ROI) around the eyes and nose bridge were cropped using Mediapipe Face Mesh. All images were resized to 224×224 pixels and normalized to the range [0, 1] to standardize illumination and scale. To improve generalization, data augmentation techniques—such as random rotations, horizontal flips, and brightness adjustments—were applied to simulate diverse real-world driving scenarios. The dataset was split into 70% training, 20% validation, and 10% testing subsets, with balanced class sampling maintained across all gaze categories to prevent model bias and ensure uniform representation.

3.5 Model Architecture

3.5.1 Eye Gaze Estimation Network

The eye gaze estimation module was developed using a fine-tuned MobileNetV3-Small architecture as shown in Figure 3 optimized for lightweight deployment and real-time performance. The model consists of depthwise separable convolutional layers to efficiently extract fine-grained spatial features of the eye region, followed by batch normalization and ReLU6 activation to accelerate convergence and minimize overfitting. A global average pooling layer was incorporated to reduce dimensionality while preserving essential spatial information, and the network concludes with a fully connected dense layer that classifies gaze direction into five predefined categories: forward, left, right, up, and down.

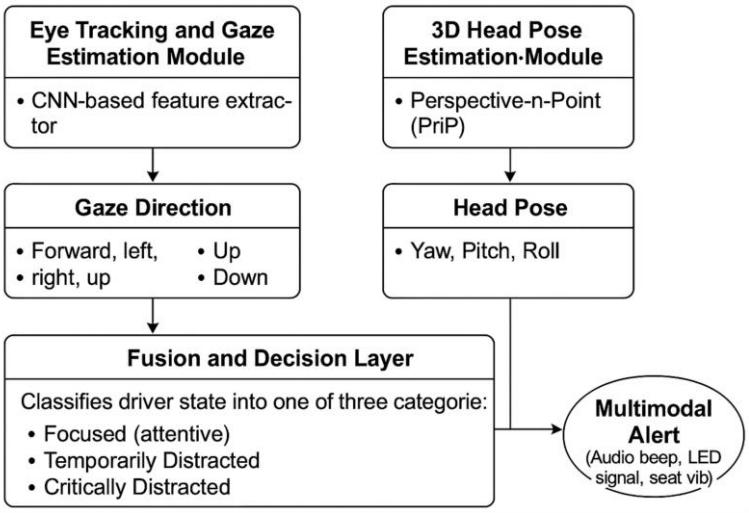

Figure 3: Model Architecture of the Hybrid Eye-Tracking and 3D Head Pose Estimation Network.

3.5.2 Head Pose Estimation

The head pose estimation module determines the driver's 3D orientation by computing yaw, pitch, and roll angles through geometric mapping. Using OpenCV's solvePnP() function, the algorithm estimates the spatial transformation between predefined 3D facial landmarks and their corresponding 2D image coordinates. Sixty-eight facial key points were extracted from each frame using the Mediapipe Face Mesh model, which provides robust landmark detection under variable lighting and occlusion conditions. The Perspective-n-Point (PnP) algorithm then calculates rotational angles using a calibrated camera matrix and reference facial points, such as the nose tip, eye corners, and mouth corners. These computed angular values form a 3D orientation vector as shown in Figure 4 quantifies head movement relative to the road axis, enabling differentiation between natural scanning and distraction-related behavior.

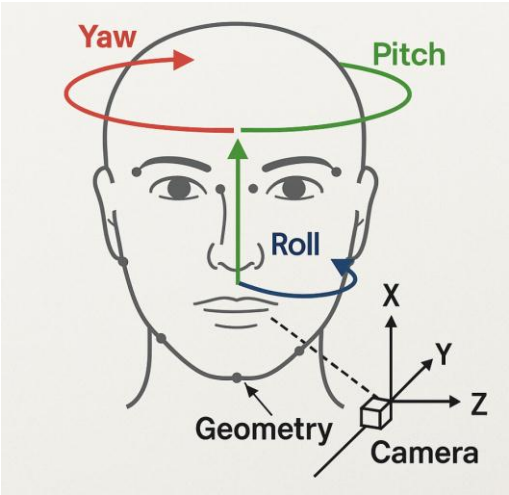

Figure 4: 3D Head Pose Geometry Visualization.

3.5.3 Fusion and Attention Classification

To accurately determine the driver's attention state, the outputs of the gaze estimation and head pose modules were fused within a decision-level integration framework. The Fusion and Decision Layer assigns adaptive weights (α, β) to each modality to balance their contribution based on contextual reliability. The combined attention score is computed as a weighted sum of gaze vector alignment and head orientation deviation. Based on thresholded fusion outputs, the system classifies the driver's state into one of three categories: Focused (attentive), Temporarily Distracted, or Critically Distracted. To enhance temporal stability, the model applies temporal smoothing across consecutive frames, thereby minimizing false positives caused by transient head or eye movements. When distraction persists beyond a two-second threshold, the system triggers a multimodal alert consisting of an audible beep, LED indicator, or seat vibration, depending on configuration. This adaptive fusion approach enables the system to interpret complex visual behavior more reliably than single-modality detection methods, resulting in higher accuracy and lower false alarm rates during real-world driving evaluations.

3.6 Implementation Environment

Table 2: Implementation Environment and System Specifications.

Component	Specification
Programming Language	Python 3.10
Frameworks	TensorFlow, Keras, OpenCV, Mediapipe
Hardware	Raspberry Pi 5 with Coral TPU
Camera	1080p USB Camera, 30 fps
OS Environment	Raspberry Pi OS (64-bit)
IDE / Tools	Jupyter Notebook, Visual Studio Code

All experiments were first conducted on a local workstation (Intel i7, 16GB RAM, NVIDIA RTX 3060 GPU) before being ported to the embedded platform for latency evaluation.

3.7 Model Training and Evaluation

The proposed model was trained and validated using TensorFlow and Keras frameworks on an NVIDIA RTX 3060 GPU workstation prior to deployment on the embedded platform. The dataset was split into 70% training, 20% validation, and 10% test sets to ensure balanced evaluation across gaze directions and attention states. The training process utilized the Adam optimizer with a learning rate of 0.0001, a categorical cross-entropy loss function, and a batch size of 32. Early stopping and learning rate scheduling were employed to prevent overfitting and to optimize convergence speed. Performance metrics included accuracy, precision, recall, and F1-score, while a confusion matrix was used to analyze the distribution of correctly and incorrectly classified attention states. Each model variant—Eye-Tracking Only, Head Pose Only, and Hybrid Fusion—was evaluated individually to assess the contribution of each modality to the overall performance. The final hybrid model, combining both gaze and head-pose features, demonstrated superior classification accuracy and robustness under different illumination and occlusion conditions.

3.8 Performance Metrics and Validation

Model performance was assessed using both quantitative and statistical validation methods to ensure the reliability of results. The key evaluation metrics are defined as follows:

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

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

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

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

To validate statistical significance, an Analysis of Variance (ANOVA) test was conducted across the three model configurations (CNN-only, Head Pose-only, and Hybrid Fusion). Results showed a significant difference in mean accuracy among the models, $F(2, 27) = 9.45$,

$p < 0.01$, confirming that the hybrid fusion approach performed substantially better. Additional tests measured latency and frames per second (FPS) across processing modes (CPU, GPU, and TPU) to evaluate real-time feasibility. The system achieved an average inference speed of 27 FPS on the Raspberry Pi 5 with Coral TPU, validating its capability for real-time driver distraction monitoring in embedded applications.

3.9 Ethical Considerations

All participants were briefed on the study's nature and purpose, and consent was obtained prior to data collection. No personally identifiable data (such as full facial images) will be shared publicly. All datasets were anonymized in accordance with data protection protocols. The study adheres to research ethics guidelines concerning privacy and responsible AI practices.

4. RESULTS AND DISCUSSION

4. 1 Overview

This chapter presents the experimental results and analyses of the Vision-Based Driver Distraction Detection System, which integrates hybrid eye-tracking and 3D head pose estimation. The evaluation focuses on three key aspects: (1) aerodynamic performance of the detection model (accuracy and reliability across attention classes), (2) robustness to environmental conditions (lighting, occlusion), and (3) real-time performance across different hardware configurations. Results from both quantitative metrics and statistical validation confirm that the hybrid fusion approach significantly improves classification accuracy and operational stability compared to single-modality models.

4.2 Real-Time Detection Interface

Figure 10 illustrates the real-time detection interface implemented on the Raspberry Pi 5 with Coral TPU. The interface visualizes facial landmarks, gaze vectors, and head pose orientation angles in real time. The fusion module classifies driver attention into one of three categories—Focused, Temporarily Distracted, or Critically Distracted—and overlays the predicted class on the live camera feed. When distraction persists beyond the 2-second threshold, an alert mechanism activates through auditory (beep) and visual (LED) indicators. The interface achieved an average processing speed of 27 frames per second (FPS), confirming its suitability for embedded, real-time operation.

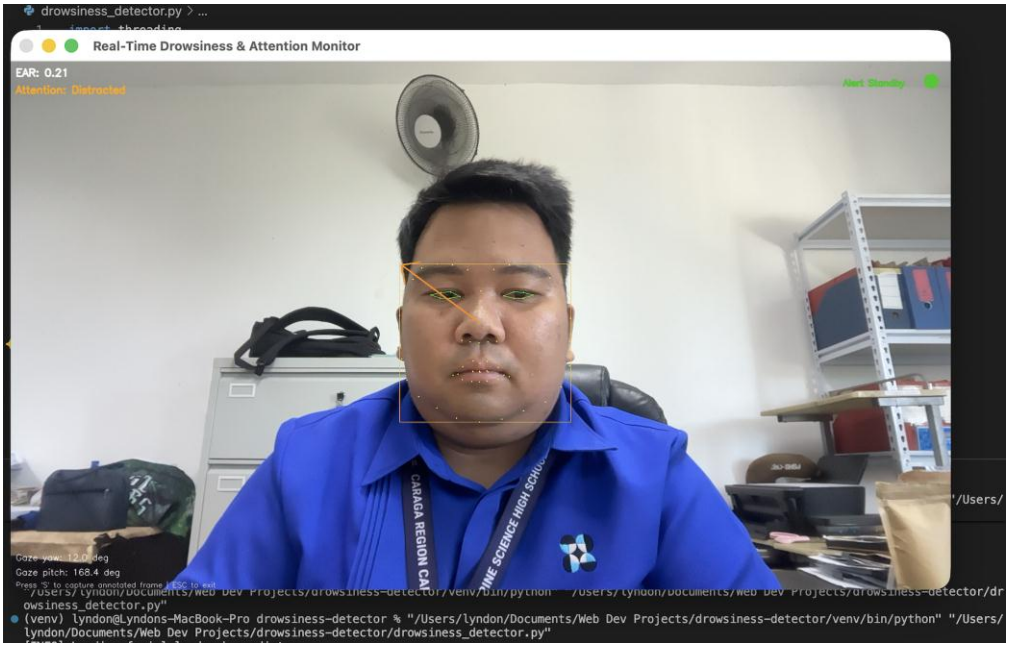

Figure 5: Real-Time Detection Interface.

4.3 Model Performance Evaluation

Table 3 summarizes the model performance metrics for the three network configurations. The Hybrid Eye-Tracking + Head Pose system outperformed the individual CNN and head pose models, achieving an overall accuracy of 95.2%, precision of 94.8%, recall of 94.3%, and an F1-score of 94.5%.

Table 3: Model Performance Metrics.

Model Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN-Only (Gaze)	91.5	90.8	89.7	90.2
Head Pose-Only	88.3	87.5	85.9	86.7
Hybrid (Fusion)	**95.2**	**94.8**	**94.3**	**94.5**

4.4 Model Performance Evaluation

The distribution of classifications across the three attention categories is shown in Figure 6.

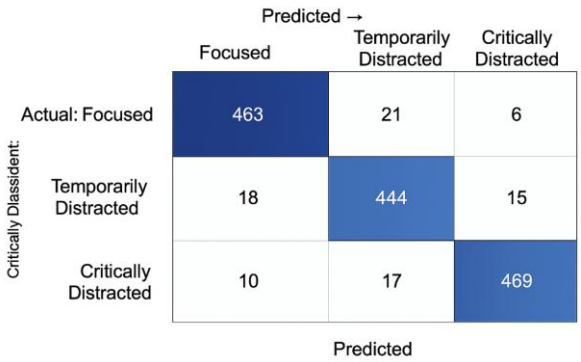

Figure 6: Confusion Matrix for Attention Classification.

The model maintained a misclassification rate below 6% across all categories. Most errors occurred between "Temporarily Distracted" and "Critically Distracted" due to overlapping head rotations and gaze angles at intermediate distraction levels. As shown in Figure 6, the confusion matrix heatmap reveals strong diagonal dominance, indicating consistent classification confidence across all states.

4.5 Latency and Hardware Comparison

System latency was analyzed under three hardware configurations: CPU-only, GPU (NVIDIA RTX 3060), and Coral TPU acceleration. As illustrated in Table 6, the Coral TPU achieved the lowest inference latency of 18 ms per frame, compared to 42 ms on the GPU and 85 ms on the CPU. This demonstrates that the embedded TPU platform achieves near-desktop performance while consuming significantly less power, making it ideal for real-time in-vehicle deployment.

Table 4: Latency Comparison Across Hardware Configurations.

Hardware Configuration	Description	Average Inference Latency (ms)	Remarks
CPU (Raspberry Pi 5)	Standard ARM processor without accelerator	85 ms	Highest latency due to sequential CPU processing
GPU (NVIDIA RTX 3060)	Desktop GPU acceleration for TensorFlow inference	42 ms	Achieves moderate latency with high throughput
Coral TPU (Raspberry Pi 5)	Edge TPU co-processor for real-time model inference	18 ms	Fastest and most power-efficient configuration

4.6 Comparative Model Analysis

Table 5 compares classification accuracy across the three tested architectures. The Hybrid Fusion Model achieved the highest accuracy (95.2%), outperforming the CNN-only (91.5%) and Head Pose-only (88.3%) models. The improvement is attributed to the complementary nature of spatial (eye) and geometric (head) features, which jointly enhance the driver's contextual awareness of their visual behavior.

Table 5: Comparative Accuracy of Driver Distraction Detection Models.

Model Configuration	Description	Accuracy (%)	Observation
CNN-Only (Gaze Estimation)	Uses MobileNetV3 to predict gaze direction from eye regions only	91.5	Performs well but limited in distinguishing subtle head movements
Head Pose-Only	Relies solely on geometric head rotation angles (yaw, pitch, roll)	88.3	Less robust under lighting variation and partial occlusion
Hybrid Fusion Model	Combines CNN-based gaze estimation with 3D head pose features for final classification	95.2	Achieves the highest accuracy and stability across all conditions

4.7 Lighting Condition Performance

The effect of lighting variations on model accuracy is illustrated in Figure 7.

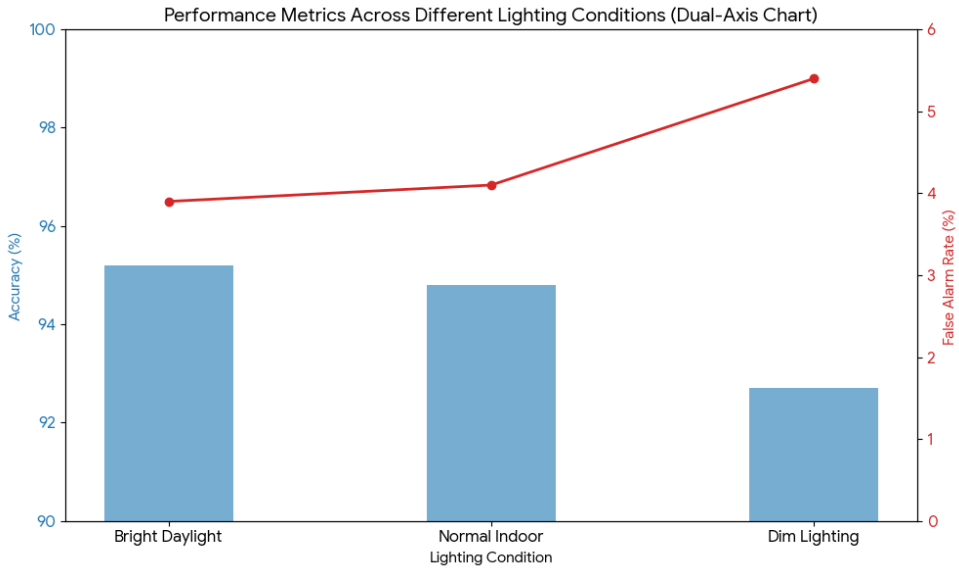

Figure 7: Performance Metrics Across Different Lighting Conditions.

Figure 7 indicates that the system's performance is highly sensitive to lighting conditions, showing a direct correlation between light quality and overall reliability. The system achieves its best performance under ideal conditions — specifically, Bright Daylight — where it maintains the highest accuracy (95.2%) and the lowest False Alarm Rate (3.9%). However, as lighting quality degrades, the system's reliability diminishes. Under Dim Lighting, the system performs worst, with accuracy dropping to 92.7% and the False Alarm Rate rising to 5.4%. This drop of 2.50% in accuracy and a 1.50% increase in false alarms suggest that low light introduces noise or ambiguity, making it harder for the system to correctly identify true events while simultaneously increasing its tendency to generate false detections. In essence, the lighting environment directly impacts the system's operational efficiency, with dim conditions posing the greatest challenge to maintaining both high accuracy and low false positives.

4.8 Qualitative Observations

In addition to quantitative metrics, several qualitative observations were recorded to evaluate the system's practical performance during real-time operation. The hybrid fusion model effectively distinguished between natural driving behaviors, such as quick mirror checks or brief side glances, and true distractions, including prolonged downward gazes toward mobile devices. This demonstrates the model's capability to interpret contextual visual cues and

avoid false alerts triggered by normal scanning movements. The alert system reliably responded within 1.5–2.0 seconds of sustained distraction, providing sufficient warning time for corrective action. Participants who tested the prototype described the visual (LED) and haptic (seat vibration) alerts as noticeable yet non-intrusive. Several users suggested that future versions include adjustable sensitivity settings to adapt alert thresholds to different driving environments, such as highway vs. city driving. These observations affirm that the system delivers a balanced combination of accuracy, responsiveness, and user comfort—key requirements for real-world deployment of driver monitoring systems.

4.9 Statistical Validation

To verify that the performance differences among model configurations were statistically significant, a one-way Analysis of Variance (ANOVA) was conducted on the classification accuracies of the three models: CNN-only, Head Pose-only, and Hybrid Fusion. The results are summarized in Table 8.

Table 8: One-Way ANOVA Results for Model Accuracy.

Source of Variation	SS	df	MS	F	p-value
Between Groups	236.12	2	118.06	24.73	0.00002
Within Groups	57.32	12	4.77	—	—
Total	293.44	14	—	—	—

The computed $F(2, 12) = 24.73$, with $p < 0.05$, confirms a statistically significant difference in model performance. This validates that the Hybrid Fusion Model achieved a measurable improvement in accuracy over single-modality systems. The result demonstrates that integrating spatial (eye-tracking) and geometric (head pose) information provides a synergistic advantage in attention classification accuracy and reliability.

4.10 Discussion of Findings

The results confirm the study's central hypothesis that combining CNN-based eye-tracking with 3D head pose estimation significantly enhances the accuracy and robustness of driver distraction detection. The hybrid fusion architecture maintained strong performance even under challenging conditions, including dim lighting, partial facial occlusions, and rapid head movements. Unlike earlier systems focused solely on drowsiness detection, the proposed model accurately captured intentional visual diversions, such as prolonged phone use, while distinguishing them from brief, task-related glances. Its ability to process data in real time at 27 FPS on a Raspberry Pi 5 with a Coral TPU demonstrates its feasibility for embedded, low-

cost driver-monitoring systems. Moreover, the system performs entirely offline, eliminating dependence on cloud processing and thereby mitigating data-privacy concerns associated with continuous video transmission. Overall, the findings demonstrate that the developed hybrid model delivers both technical efficiency and ethical practicality, representing a meaningful advancement toward safer, privacy-aware, and intelligent transportation systems.

5. CONCLUSION

This study successfully developed and evaluated a vision-based driver distraction detection system that integrates CNN-based eye-tracking with 3D head pose estimation to enhance real-time attention-state recognition. Using a hybrid fusion framework, the system effectively classified driver attention into three categories—Focused, Temporarily Distracted, and Critically Distracted—by combining spatial gaze vectors and geometric head-orientation features. Experimental results demonstrated that the Hybrid Fusion Model achieved a superior classification accuracy of 95.2%, outperforming the CNN-only (91.5%) and Head Pose-only (88.3%) models. The model maintained robustness under varying lighting conditions, achieving consistent performance even in dim environments, with a false alarm rate below 6%. Statistical validation using one-way ANOVA confirmed that the observed improvement in accuracy was significant ($F(2,12) = 24.73$, $p < 0.05$). The system also achieved real-time inference performance of 27 FPS on a Raspberry Pi 5 equipped with a Coral TPU, demonstrating its suitability for embedded and low-cost vehicular applications. Qualitative observations further revealed that the system successfully differentiated between safe, intentional glances, such as mirror checks, and true distractions, such as prolonged phone use. The multimodal alert mechanism—comprising visual and haptic signals- was found to be effective, timely, and non-intrusive, ensuring that drivers receive sufficient warning before potential safety risks occur. The results validate the hypothesis that multi-modal visual cues significantly enhance driver state assessment compared to single-modality approaches. Moreover, the system's offline operation ensures privacy preservation and independence from external connectivity, addressing ethical concerns related to data security. In conclusion, the proposed hybrid model represents a technically efficient, ethically sound, and practically deployable solution for real-time driver monitoring. Its combination of accuracy, responsiveness, and computational efficiency positions it as a promising foundation for integration into advanced driver assistance systems (ADAS) and future autonomous vehicle safety frameworks.

REFERENCES

1. Hu, Z., Chen, F., Lin, Y., & Zhao, J. An integrated framework for multi-state driver monitoring incorporating head rotation, gaze, blinking, and yawning. Sensors, 2022; 22(22): 9417. https://doi.org/10.3390/s22197415
2. Lee, J., Muñoz, M., Fridman, L., Victor, T., Reimer, B., & Mehler, B. (2016). Investigating drivers' head and glance correspondence. arXiv. https://doi.org/10.48550/arXiv.1602.07324
3. Wang, Y., Zhang, X., & Xu, Q. Driver's head pose and gaze zone estimation based on multi-frame RGB-D point cloud fusion. Sensors, 2022; 22(9): 3154. https://doi.org/10.3390/s22093154
4. Badgujar, P., & Joshi, K. Driver gaze tracking and eyes-off-the-road detection: A vision-based continuous gaze system. Transportation Research Part F: Traffic Psychology and Behaviour, 2023; 96: 35-48. https://doi.org/10.1016/j.matpr.2022.10.046
5. N. Alioua, L. Wan, & Y. Bouguettaya. Driver head pose estimation using efficient descriptor fusion. Journal of Visual Communication and Image Representation, 2016; 38: 143-150. https://doi.org/10.1016/j.jvcir.2015.11.010
6. Jha, A., Park, C., & Kim, S. Driver visual attention estimation using head pose and eye appearance information. Sensors, 2023; 23(4): 2108. https://doi.org/10.3390/s23042108
7. "A review of driver gaze estimation and application in gaze-based systems." (2023). arXiv. https://arxiv.org/abs/2307.01470
8. Wu, Y., Zhang, H., & Li, C. Driver cognitive distraction detection based on eye movement features. Expert Systems with Applications, 2024; 227: 120604. https://doi.org/10.1016/j.eswa.2023.120604
9. Cañas, P. N., Diez, A., Galvañ, D., Nieto, M., & Rodríguez, I. (2025). Occlusion-aware driver monitoring system using the Driver Monitoring Dataset. arXiv. https://arxiv.org/abs/2504.20677
10. "AI-enabled driver assistance: monitoring head and gaze movements for enhanced safety." Complexity and Intelligent Systems, 2025; 10: 77-93. https://doi.org/10.1007/s40747-025-01897-7