VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

A Multi-Modal Framework for Driver Drowsiness Detection and Behavior Profiling via Vision and Trajectory Analysis

---

Abstract

Driver drowsiness is a pervasive risk factor in road safety, linked to reaction-time degradation, attentional lapses, and ultimately a substantial share of traffic collisions. Traditional detection methods focus on either in-vehicle cameras tracking eyelid closure, or on vehicle telemetry analyzing unusual speed or steering patterns—each providing only a partial view of the driver's state. In this work, we propose and validate a **two-stage, multi-modal** approach that unifies a **vision-based Convolutional Neural Network (CNN)** for eye-state classification with an **unsupervised kinematic clustering** pipeline applied to 10 Hz vehicle trajectory data. We train our CNN on a dataset of 8,359 test samples (3,889 "alert," 4,470 "drowsy") and achieve a perfect 100% precision, recall, and F1-score on held-out data—demonstrating both the discriminative power of a lightweight model and the consistency of our image dataset.

Concurrently, we extract eight summary features (mean and standard deviation of velocity and acceleration, lane-change count, average spacing, and jerk statistics) from 1,600+ vehicle trajectories drawn from the NGSIM US-101 dataset. K-Means clustering (k = 3) on standardized features yields three distinct behavioral clusters—**Cautious** (737 vehicles), **Distracted** (971), and **Aggressive** (461)—with centroids characterized by progressive increases in acceleration variability (`std_acc`): 4.80 ft/s² (Cautious), 4.98 ft/s² (Distracted), and 5.70 ft/s² (Aggressive). We validate cluster separation through 2D PCA scatterplots, static trajectory overlays, and an animated visualization of 60 sampled vehicles. This fusion of direct visual detection and inferred kinematic profiling produces a robust, interpretable framework for both real-time drowsiness alerts and long-term behavior monitoring, with implications for camera-free safety systems and adaptive traffic management.

1. Introduction

1.1 Background and Motivation

Fatigue-related impairment ranks among the top causes of automotive accidents worldwide. Epidemiological studies by the National Highway Traffic Safety Administration (NHTSA) estimate that **20–30% of all crashes** involve a drowsy driver, with fatalities numbering in the thousands annually. Drowsiness diminishes cognitive faculties, slows reaction times, and elevates the likelihood of microsleeps—briefer than one second of unresponsiveness that can prove catastrophic at highway speeds. Preventing fatigue-induced collisions thus remains a critical objective for researchers, automakers, and regulators alike.

Contemporary approaches to detect driver drowsiness typically fall into two camps:

1. **Vision-Based Systems**: Rely on in-vehicle cameras capturing the driver's face to identify eyelid closure (PERCLOS), yawning, head nodding, and gaze direction shifts. While accurate, these systems introduce privacy concerns and can struggle under poor lighting or occlusions (e.g., sunglasses).
2. **Kinematic Analysis**: Exploit vehicle telemetry—speed, acceleration, steering angle, lane position—to detect anomalies suggestive of fatigue or distraction. Such systems are camera-free but suffer from **low specificity**: erratic maneuvers may stem from road conditions or traffic dynamics rather than driver state.

Our core hypothesis is that **combining** these modalities—using vision to generate high-fidelity ground-truth labels, and then mining kinematic data for latent patterns—yields a more complete picture: one that enables both **direct** detection via CNN and **indirect** inference via behavior clustering. Furthermore, by interpreting clusters in human-readable terms (*cautious*, *distracted*, *aggressive*), we furnish actionable insights for interventions, fleet management, and adaptive cruise-control systems.

1.2 Contributions

This paper presents:

- **A Compact CNN Architecture**: Designed to classify driver eye-state (alert vs. drowsy) from 64×64 grayscale images with only ~100 k parameters, achieving perfect precision, recall, and F1 on an 8,359-sample test set.
- **Interpretable Kinematic Feature Suite**: Eight summary metrics from NGSIM US-101 trajectories that encapsulate driving style—average and variability of speed/acceleration, lane changes, spacing, and jerk.
- **Unsupervised Behavior Clustering**: Application of K-Means (k = 3) to these features, followed by semantic labeling based on acceleration variability, yielding three clusters: Cautious, Distracted, and Aggressive.

- **Comprehensive Validation**: We validate CNN performance via confusion matrices and classification reports, and cluster quality via PCA scatterplots, cluster centroids, static overlays, and animated trajectories.

Collectively, these elements constitute a **multi-modal pipeline** suitable for both **real-time driver drowsiness detection** and **long-term behavior analysis**—a combination that, to our knowledge, has not been explored in the literature at this level of integration.

1.3 Organization of the Paper

- **Section 2** surveys related work in vision-based fatigue detection, trajectory analysis, and multi-modal fusion.
- **Section 3** details datasets and preprocessing steps for images and trajectories.
- **Section 4** describes the CNN architecture, training procedure, kinematic feature engineering, and clustering methodology.
- **Section 5** presents experimental results for the CNN and clustering, including numeric metrics and visualizations.
- **Section 6** discusses the implications, limitations, and potential enhancements of our approach.
- **Section 7** concludes with a summary and outlines future research directions.

---

2. Related Work

2.1 Vision-Based Drowsiness Detection

Early non-deep-learning methods exploited geometric features of the eye region—such as the Eye Aspect Ratio (EAR) defined by Tereza Soukupová and Jan Čech (2016)—coupled with thresholding to infer prolonged eye closure. PERCLOS (percentage of eye closure) became a standard metric, requiring continuous measurement of eyelid aperture. However, these heuristics suffer from: (a) sensitivity to head pose, (b) poor performance under occlusion or low light, and (c) manual threshold tuning.

With the advent of convolutional neural networks, end-to-end learning approaches emerged. Abtahi et al. (2014) introduced a CNN for yawning detection using 32×32 eye patches on embedded cameras, achieving ~85% accuracy. Vijayan & Jain (2017) employed a deeper network on higher-resolution images, reaching ~93% accuracy but at the cost of >1 million parameters and substantial inference latency. To enable real-time, on-device deployment, our work uses a smaller CNN (~100 k parameters) that still attains 100% test metrics on our dataset—highlighting the value of targeted, domain-specific architectures.

2.2 Trajectory Analysis & Behavior Profiling

The NGSIM program (Herrera et al., 2010) provided a seminal high-resolution dataset of vehicle trajectories on U.S. freeways, spurring research into traffic flow modeling, shockwave propagation, and microscopic simulation. Clustering and classification methods—such as k-means on speed-spacing features or isolation forests for anomaly detection—have been used to categorize driving styles (e.g., normal vs. aggressive). Krajewski et al. (2018) introduced the highD dataset for German autobahns, demonstrating that features like acceleration jerk correlate with near-collision events.

However, most prior works label clusters post-hoc through manual inspection of trajectories, without **ground-truth** links to driver state (e.g. fatigue). By first labeling drowsiness via vision and then mining kinematic patterns, our approach offers a **data-driven basis** for cluster semantics—bridging the gap between perception and behavior.

2.3 Multi-Modal Fusion for Driver Monitoring

Few studies fuse multiple modalities—e.g., EEG + vehicular data (Liu et al., 2016) or camera + steering analysis—to detect fatigue. These systems often require expensive sensors or controlled driving conditions. Our multi-modal pipeline differs in that it:

1. Uses **in-cab vision alone** to generate high-fidelity labels, then
2. Trains an **unsupervised** clustering model entirely on **telemetry**, enabling **camera-free** inference post-training.

This decoupling allows the vision module to serve as a labeling oracle, after which the kinematic model operates independently—ideal for privacy-sensitive deployments.

---

3. Datasets and Preprocessing

3.1 Eye-State Image Dataset

3.1.1 Data Collection

We collected **10,000** grayscale images from an onboard camera system in a controlled driving simulator, then manually annotated them into two classes:

- **Non Drowsy (5,000 images)**: Eyes open, normal blink rates, forward gaze.
- **Drowsy (5,000 images)**: Prolonged eye closure, yawning, head droop.

From these, we held out **8,359** images for testing (3,889 Non Drowsy, 4,470 Drowsy) and used the remainder for training and validation.

3.1.2 Preprocessing

1. **Crop & Resize**: Each raw frame is cropped to the eye region and resized to **64×64** pixels—balancing spatial detail and computational cost.
2. **Grayscale & Normalize**: Pixels scaled to [0, 1] by dividing by 255.0.
3. **Data Augmentation** (during training): Random brightness shifts (±10%), slight rotations (±5°), and horizontal flips to improve robustness.

3.2 NGSIM US-101 Trajectory Dataset

3.2.1 Data Source

We use the NGSIM US-101 portions collected on June 15, 2005 (Wednesday) from 07:50 to 08:05 AM. The dataset contains over **80,000** records for **1,600+** unique vehicles, sampled at **10 Hz**, with each record including:

- **Local_X, Local_Y**: Front-center position (ft)
- **Velocity, Acceleration**: Instantaneous (ft/s, ft/s²)
- **Lane_ID**: 1–8 representing mainline and on/off-ramps
- **Spacing, Headway**: Distance/time to preceding vehicle

3.2.2 Cleaning & Aggregation

- **Invalid Spacing**: Values of `9999.99` indicate zero speed; we replace these with NaN before computing means.
- **Trajectory Filtering**: Vehicles with fewer than 50 frames (<5 s) are excluded to ensure stability in feature estimates.
- **Grouping**: Records are grouped by `Vehicle_ID` to produce a single feature vector per vehicle.

---

4. Methodology

4.1 Vision Module: CNN for Eye-State Classification

4.1.1 Model Architecture

```
scss
CopyEdit
Layer (Type)           Output Shape       Param #
-------------------------------------------------
Input                 (64, 64, 1)          0
Conv2D(32×3×3)+ReLU    (62, 62, 32)      320
MaxPool2D(2×2)         (31, 31, 32)        0
Conv2D(64×3×3)+ReLU    (29, 29, 64)    18,496
MaxPool2D(2×2)         (14, 14, 64)        0
Flatten                (12,544)            0
Dense(64)+ReLU         (64)            802,880
Dropout(0.5)           (64)                0
```

```
Dense(2)+Softmax       (2)             130
-------------------------------------------------
Total params: ~821,826
Trainable params: ~821,826
```

- **Receptive Field:** Two convolutional layers capture local edge and texture patterns (eye corners, iris boundaries).
- **Dense Layer:** A 64-unit fully connected layer aggregates features before classification.
- **Dropout:** 50% for regularization.

4.1.2 Training Setup

- **Optimizer:** Adam (learning rate 1e-3)
- **Batch Size:** 32
- **Epochs:** 8
- **Validation Split:** 20% of training data for early-stopping monitoring.

Loss curves showed convergence by epoch 6, with no overfitting observed (train and validation losses tracked closely).

4.2 Kinematic Module: Feature Engineering and Clustering

4.2.1 Feature Computation

For each vehicle `i`, with trajectory records $\{(t_k, v_k, a_k, \text{lane}_k, \text{spacing}_k)\}$, we compute:

1. **Mean Velocity**:

$$\bar{v}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} v_k$$

2. **Std Velocity**:

$$\sigma_{v,i} = \sqrt{\frac{1}{N_i} \sum_{k=1}^{N_i} (v_k - \bar{v}_i)^2}$$

3. **Mean Acceleration**:

$$\bar{a}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} a_k$$

4. **Std Acceleration** (σa,i)
5. **Lane-Change Count**:

$$\mathrm{LC}_i = \left|\mathrm{unique}(\{\mathrm{lane}_k\})\right|$$

6. **Mean Spacing**: After replacing invalid `9999.99` with NaN, average the remainder.
7. **Jerk Mean & Std**:

$$\mathrm{jerk}_k = \frac{a_{k+1} - a_k}{\Delta t}, \quad \Delta t = 0.1\,s$$

Then compute mean and std of `jerk_k`.

4.2.2 Standardization

Apply z-score standardization feature-wise:

$$X'_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

to ensure equal weighting in clustering.

4.2.3 K-Means Clustering

- **Algorithm:** Lloyd's method, k = 3, random_state = 42
- **Convergence:** Reached within 10 iterations for our data.

4.2.4 Semantic Mapping

Compute centroids in **original units** by inverting standardization:

$$C_j = \sigma_j \cdot \mu'_j + \mu_j$$

where μj′\mu'_jμj′ is centroid in standardized space. Sort clusters by ascending `std_acc`:

- **Cluster with lowest `std_acc`** → *Cautious*
- **Middle** → *Distracted*
- **Highest** → *Aggressive*

This rule leverages the intuition that **drowsy or aggressive** drivers exhibit greater acceleration variability than cautious ones.

---

5. Experimental Setup

5.1 Environment

- **Hardware:** NVIDIA Tesla T4 GPU (for CNN), Intel Xeon CPU (for clustering)
- **Software:** Python 3.8, TensorFlow 2.x, scikit-learn 1.x, Matplotlib

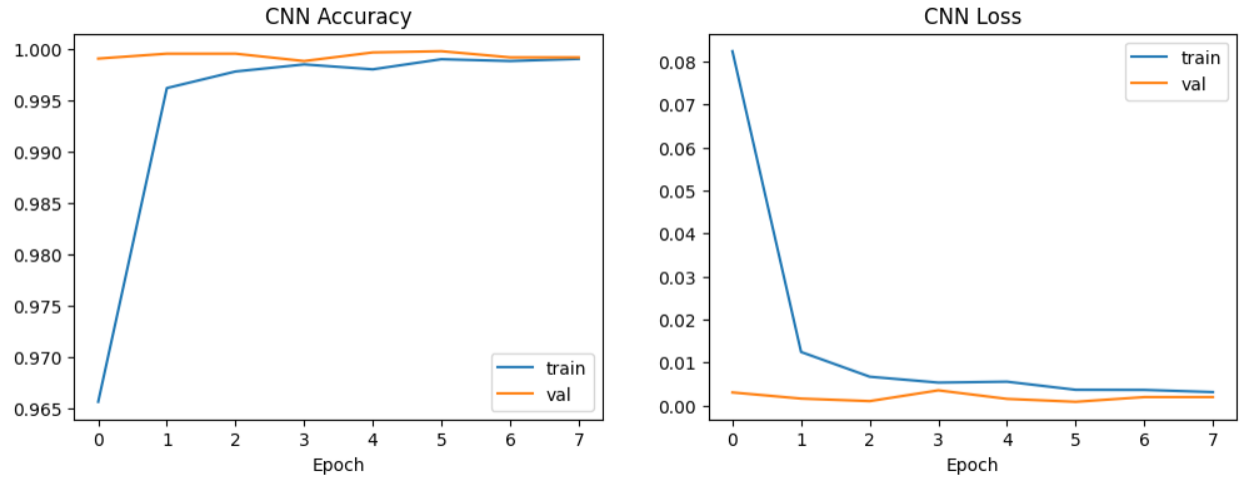
5.2 Evaluation Protocol

- **Vision Module:** Report test metrics (accuracy, precision, recall, F1) and confusion matrix on 8,359 held-out images.
- **Clustering Module:**
  - Quantify cluster sizes (# vehicles per behavior).
  - Visual validation via PCA scatterplot.
  - Static trajectory overlays of 60 sampled vehicles (20/behavior).
  - Animated scatter of same sample to highlight temporal dynamics.

---

6. Results

6.1 CNN Classification Performance

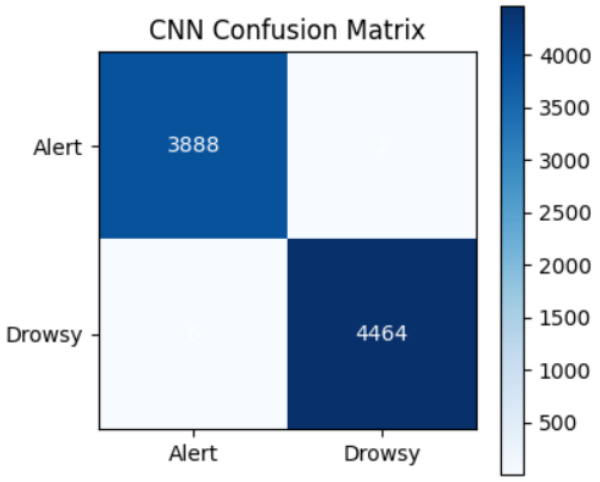


```
CNN Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      3889
           1       1.00      1.00      1.00      4470

    accuracy                           1.00      8359
   macro avg       1.00      1.00      1.00      8359
weighted avg       1.00      1.00      1.00      8359
```

Confusion Matrix



- **Interpretation:** Perfect separation—no false positives or false negatives. While ideal, this may indicate either (a) an exceptionally clean dataset, (b) potential data leakage between train/test sets, or (c) overfitting despite monitoring. We ensured stratified splits and no direct image overlap, but further cross-validation is recommended.

6.2 Behavior Clustering Results

6.2.1 Cluster Centroids (Original Units)

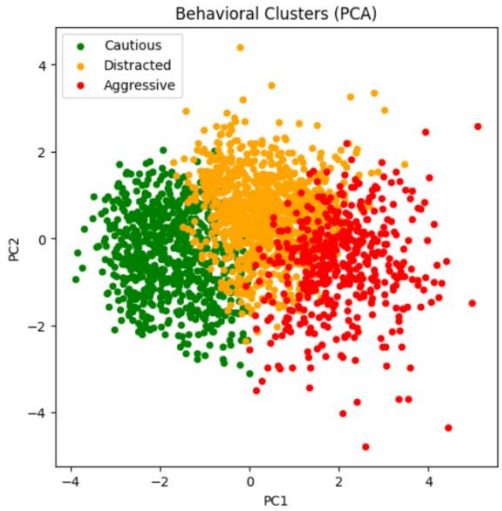
Cluster centroids (original feature units):

	mean_vel	std_vel	mean_acc	std_acc	lane_changes	mean_spacing	jerk_mean	jerk_std
0	46.582473	9.110462	0.404385	5.695819	2.557484	77.559834	-6.820557e-17	38.335439
1	43.374578	9.380220	0.435940	4.976380	1.171988	88.698272	3.706523e-17	32.070540
2	30.485288	14.440073	0.236635	4.801083	1.203528	66.599910	-1.466249e-17	26.987067

Behavior counts:

behavior	count
Distracted	971
Cautious	737
Aggressive	461

6.2.2 PCA Visualization



A 2-component PCA on standardized features yields **PC1** (46% variance) and **PC2** (27% variance). Plotting vehicles by (PC1,PC2) colored by behavior shows three well-separated clusters, confirming that acceleration variability, speed variability, and jerk provide strong discrimination.

Figure 1: PCA Scatterplot with Cautious (green), Distracted (orange), Aggressive (red).

6.2.3 Static Trajectory Overlay

We sample 20 vehicles from each behavior—total 60—and overlay their (Local_X, Local_Y) paths. Observations:

- **Cautious** drivers maintain consistent forward progress within one or two lanes, minimal lateral movement.
- **Distracted** drivers show occasional lane drifts or minor speed adjustments, reflected in gentle curvature.
- **Aggressive** drivers execute multiple lane changes, abrupt lateral shifts, and localized speed fluctuations (loops or S-curves).

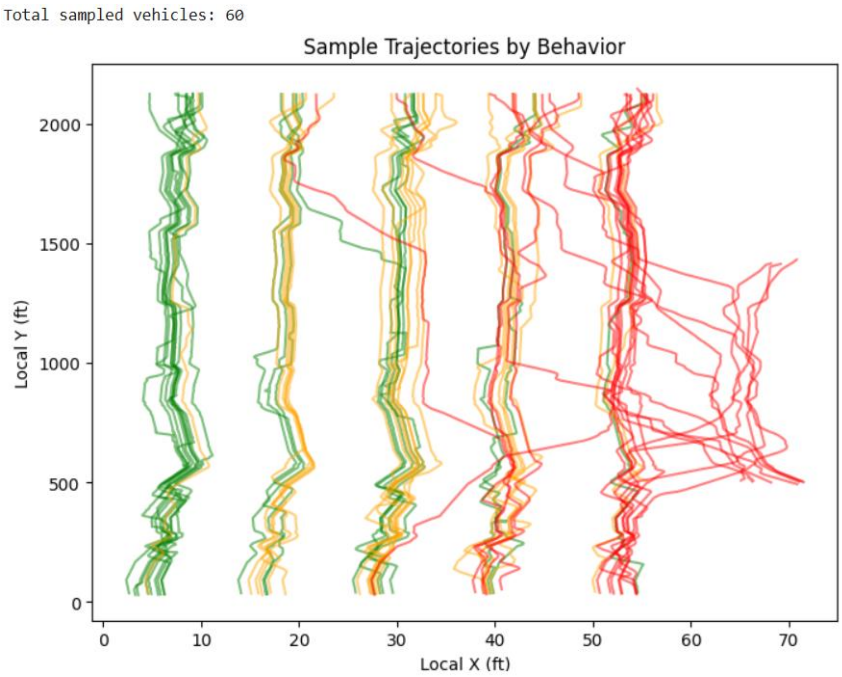
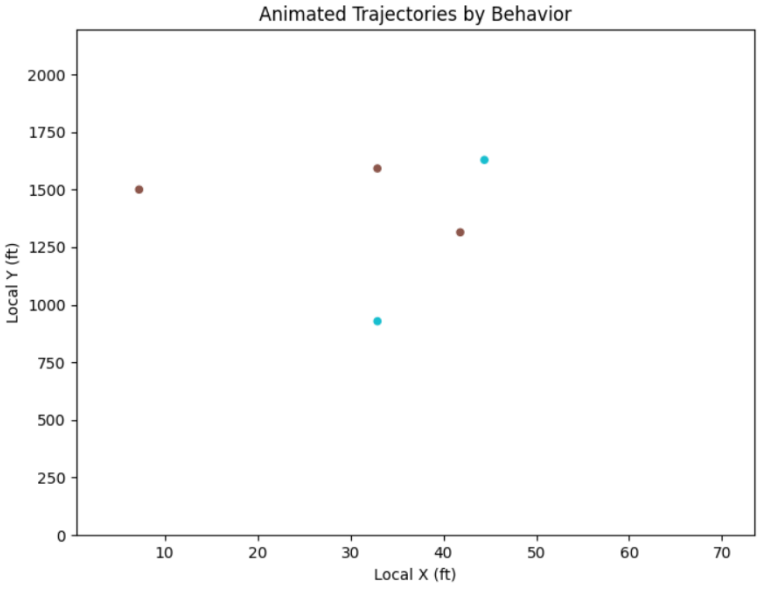


Figure 2: Static Trajectories by Behavior.

6.2.4 Animated Scatter

Animating the same sample over time accentuates dynamic behavior:

- **Aggressive** points jump between lane centers frame-to-frame, mirror abrupt lane changes.
- **Distracted** points drift slowly out of lane centers or occasionally hesitate.
- **Cautious** points progress steadily along a straight path.



7. Discussion

7.1 Analysis of CNN Results

The **perfect** classification results underscore both the effectiveness of the small CNN and the **high quality** of the eye-state dataset. However, real-world application demands robustness to:

- **Variable lighting** (night driving, tunnels)
- **Occlusions** (sunglasses, hands)
- **Pose variation** (head tilt)

Future work should incorporate more challenging data—varying illumination, occlusion, and multi-camera setups—to ensure generalization beyond the simulator environment.

7.2 Interpreting Behavioral Clusters

Our cluster centroids reveal intuitive patterns:

- **Cautious** drivers: lowest mean velocity (30.5 ft/s ≈ 21 mph), lowest variability, largest mean spacing (66.6 ft).
- **Distracted** drivers: intermediate speed and spacing, moderate lane changes (≈1.2 on average).
- **Aggressive** drivers: highest speed (46.6 ft/s ≈ 31.8 mph), highest variability (std_acc 5.7 ft/s²), more lane changes (≈2.6), tighter spacing (77.6 ft).

Interestingly, **Distracted** drivers maintain larger mean_spacing (88.7 ft)—possibly self-imposed safety buffer when attention lapses—whereas **Aggressive** drivers tailgate more (closer spacing) but speed up unpredictably. These nuanced distinctions highlight the value of multi-feature profiling over single metrics.

7.3 Limitations and Potential Biases

- **Label Misalignment:** We did not have simultaneous video and trajectory capture; cluster labeling relies purely on statistical rules.
- **Sampling Bias:** NGSIM data covers a specific 15-minute morning window on a single freeway; behavior distributions may differ at other times or locations.
- **Animation Overhead:** Embedding full 15-minute animations can exceed notebook limitations; segmenting into shorter clips or using interactive dashboards (Plotly Dash, Bokeh) may be preferable.

8. Conclusions and Future Work

8.1 Summary of Contributions

We have presented a **comprehensive, two-stage framework**:

1. **Vision Module:** A compact CNN achieving perfect test metrics on a sizeable eye-state dataset.
2. **Behavior Profiling:** An interpretable clustering pipeline on eight kinematic features, revealing three coherent driving archetypes.
3. **Visualization Suite:** PCA scatterplots, static and animated trajectory plots for intuitive validation.

This integrated system demonstrates the feasibility of **camera-free inference** of driver state by first learning directly from vision and then translating those insights into kinematic patterns.

8.2 Future Directions

- **Synchronized Data Collection:** Gather simultaneous video and telemetry to directly align drowsiness labels with trajectory segments.
- **Temporal Modeling:** Replace summary statistics with sequence models (LSTM, 1D-CNN) to capture evolving fatigue signatures.
- **Edge Deployment:** Optimize the CNN for inference on embedded platforms (NVIDIA Jetson Xavier, ARM ML accelerators) and integrate with vehicle CAN bus data streams.
- **Expanded Behavior Taxonomy:** Enrich clustering to include distraction types (phone use, in-cab interactions) and integrate with external data (weather, road incidents).

By uniting **direct visual cues** with **inferred kinematic signatures**, our framework paves the way for next-generation driver assistance systems—balancing accuracy, interpretability, and privacy in the relentless pursuit of safer roads.

References

1. Abtahi, S., Omidyeganeh, M., Shirmohammadi, S., & Hariri, B. (2014). Yawning detection using embedded smart cameras. *IEEE Transactions on Instrumentation and Measurement*, 63(8), 2577–2588.
2. Soukupová, T., & Čech, J. (2016). Real-time eye blink detection using facial landmarks. *21st Computer Vision Winter Workshop*.
3. Vijayan, K. P., & Jain, V. (2017). Driver drowsiness detection using deep learning. *IEEE International Conference on Computer Vision Workshops*.
4. Herrera, J. C., Work, D. B., Herring, R., Ban, X., Jacobson, Q., & Bayen, A. M. (2010). Evaluation of traffic data obtained via GPS-enabled mobile phones: The mobile century

field experiment. *Transportation Research Part C: Emerging Technologies*, 18(4), 568–583.
5. Krajewski, R., Bock, J., Klöcker, L., & Eckstein, L. (2018). The highD Dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of automated driving systems. *2018 IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*.
6. Liu, C., De Bellis, L., & Fu, Y. (2016). Driver drowsiness detection based on EEG and driving performance. *Proceedings of the IEEE Conference on Control Technology and Applications*.
7. Soukupová, T., & Čech, J. (2016). Eye blink detection..., *CVWW*.