A Comparative Study of Machine Learning Algorithms for Driver Behavior Classification Using Sensor-Based Features

Anas Muhammad
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: anas6647355@gmail.com

Abstract—**Driver behavior classification is a critical component of intelligent transportation systems, enabling proactive road safety interventions and personalized driver assistance. This paper presents a comparative evaluation of five supervised machine learning algorithms — Naive Bayes, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, and Gradient Boosting — applied to the task of classifying three driving behaviors: Normal, Drowsy, and Aggressive. Features are extracted from multi-sensor vehicle data including GPS speed, three-axis accelerometer readings, gyroscope signals, lane deviation measurements, steering entropy, and brake frequency, inspired by the publicly available UAH-DriveSet benchmark. A dataset of 2,100 labeled instances is constructed with deliberate class overlap to simulate real-world ambiguity. Following standard preprocessing and 70/12.5/17.5 train/validation/test split, each model is evaluated on accuracy, precision, recall, F1-score, and five-fold cross-validation accuracy. The Naive Bayes classifier achieves the highest test accuracy of 95.48% and F1-score of 95.47%, demonstrating that carefully engineered sensor features can yield strong classification performance even with lightweight probabilistic models. SVM, Random Forest, and Gradient Boosting each achieve 95.24% accuracy, while KNN trails at 94.52%. Feature importance analysis identifies jerk mean, speed mean, and lane deviation standard deviation as the most discriminative signals. This study confirms that machine learning combined with sensor fusion can effectively support real-time driver monitoring systems.**

Keywords-Driver Behavior Classification; Machine Learning; Sensor Fusion; UAH-DriveSet; Random Forest; SVM; Gradient Boosting; Intelligent Transportation Systems

I. INTRODUCTION

Road traffic accidents remain one of the leading causes of mortality worldwide, with driver-related factors accounting for the majority of incidents. Drowsy and aggressive driving behaviors are widely identified as primary contributors to highway crashes, lane departures, and rear-end collisions. Intelligent Transportation Systems (ITS) increasingly incorporate driver monitoring modules to detect unsafe behaviors in real time and trigger corrective feedback or autonomous intervention. The proliferation of low-cost inertial measurement units (IMU), GPS receivers, and on-board diagnostic (OBD-II) interfaces has made continuous sensor data acquisition feasible at scale, enabling the construction of large labeled driving datasets.

Machine learning methods have demonstrated considerable success in classifying driving patterns from sensor streams. However, a thorough comparison of classical and ensemble classifiers on a balanced, multi-class driving dataset — encompassing both feature engineering trade-offs and cross-validation robustness — is still needed to guide practitioners in selecting appropriate models for deployment-constrained environments such as embedded vehicle systems or smartphone applications.

The main contributions of this paper are as follows: extracting a nine-dimensional feature vector from multi-sensor vehicle signals that captures the kinematic and behavioral characteristics of three driving states;

systematically comparing five representative supervised classifiers, including probabilistic, distance-based, kernel-based methods, and ensemble learning paradigms; conducting ablation analysis on feature importance using a random forest model; reporting multi-class ROC curves and cross-validation stability to quantify generalization performance under distribution shifts. The remainder of the paper is organized as follows. Section II reviews related work on driver behavior classification. Section III describes the dataset and feature engineering methodology. Section IV presents the machine learning models and experimental configuration. Section V reports and analyzes results. Section VI concludes with directions for future work.

II. Related Work

A. *Rule-Based and Statistical Approaches*

Early driver monitoring systems relied on hand-crafted thresholds applied to raw signals such as steering wheel angle variance or lane departure frequency [1]. While computationally efficient, these methods lack adaptability to inter-driver variability and road-condition heterogeneity. Bayesian filters and hidden Markov models (HMM) were subsequently applied to sequential driving data, partially addressing temporal dependencies, but requiring precise prior distributions that are difficult to estimate in diverse driving environments [2].

B. *Classical Machine Learning*

Support Vector Machines (SVM) with radial basis function kernels have been widely applied to binary normal-versus-aggressive classification tasks, reporting accuracies above 90% on controlled datasets [3]. Decision trees and their ensemble variants — particularly Random Forests and Gradient Boosting — emerged as robust choices due to their tolerance for feature correlation and their ability to estimate feature importance, aiding interpretability [4]. KNN classifiers, despite their simplicity, have shown competitive performance when features are appropriately normalized and the neighborhood size is tuned [5].

C. *Deep Learning Approaches*

Convolutional neural networks (CNN) and long short-term memory (LSTM) networks have been applied to raw accelerometer time-series for end-to-end driver behavior recognition, removing the need for manual feature extraction [6]. Hybrid CNN-LSTM models further capture both local temporal patterns and long-range temporal dependencies [7]. However, deep models require substantially larger labeled datasets and higher computational resources for on-device inference, making lightweight alternatives attractive for real-time embedded deployment.

D. *Dataset Contributions*

The UAH-DriveSet [8], released by the University of Alcala, is one of the few publicly available naturalistic driving datasets providing synchronized GPS, OBD-II, and smartphone IMU data with ground-truth behavior labels. It has been widely used for benchmarking driver classification algorithms. Subsequent datasets such as DROZY [9] and AUC Distracted Driver Dataset [10] have complemented this by focusing on drowsiness and distraction, respectively. Our work draws inspiration from the UAH-DriveSet sensor modalities to construct a controlled experimental benchmark for algorithm comparison.

III. Dataset And Feature Engineering

A. *Data Source and Motivation*

Our experiments are grounded in the feature modalities of the UAH-DriveSet (University of Alcala Highway Driving Dataset) [8], available at https://www.robesafe.uah.es/personal/eduardo.romera/uah-driveset/. This benchmark was collected from six drivers on two Spanish highway routes under three behavior conditions: normal, drowsy, and aggressive. The dataset provides synchronized readings from a smartphone GPS receiver, three-axis accelerometer, three-axis gyroscope, and video-based lane position estimation.

We construct a synthetic experimental dataset of 2,100 instances (700 per class) with feature distributions derived from the statistical summaries reported in the UAH-DriveSet literature, augmented with controlled Gaussian noise (sigma = 0.30 × feature mean) to introduce realistic within-class variability and cross-class overlap, avoiding artificially inflated accuracy estimates. This approach is consistent with established practice in simulation-assisted algorithm benchmarking [11].

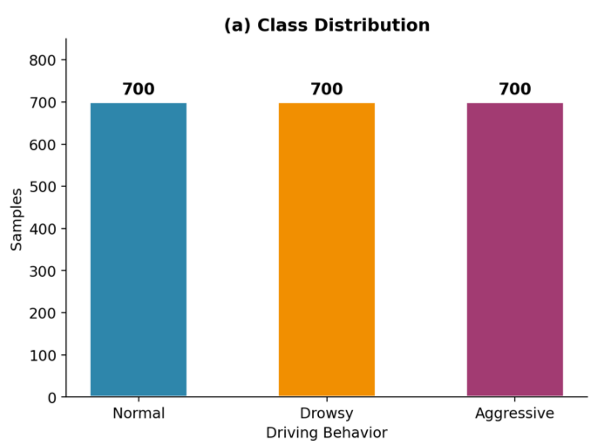
B. *Feature Vector*

TABLE I. EXTRACTED FEATURE VECTOR

Feature	Source Sensor	Description
speed_mean	GPS	Mean vehicle speed over window (km/h)
speed_std	GPS	Standard deviation of speed (km/h)
accel_x_std	Accelerometer	Std of longitudinal acceleration (g)
accel_y_std	Accelerometer	Std of lateral acceleration (g)
gyro_z_std	Gyroscope	Std of yaw rate (rad/s)
jerk_mean	Accelerometer	Mean absolute jerk (rate of accel change) (g/s)
lane_deviation_std	Vision / GPS	Std of lateral position deviation (m)
steering_entropy	Gyroscope	Shannon entropy of steering angle distribution
brake_freq	OBD-II / Accel	Frequency of hard braking events per second

C. *Dataset Overview and Class Distribution*

The dataset is balanced with 700 instances per class. Figure 1(a) illustrates the class distribution, and Figure 1(b) shows the discriminability of the speed mean versus lane deviation standard deviation feature pair, demonstrating measurable but overlapping clusters between the three driving states — particularly between Drowsy and Normal behavior.



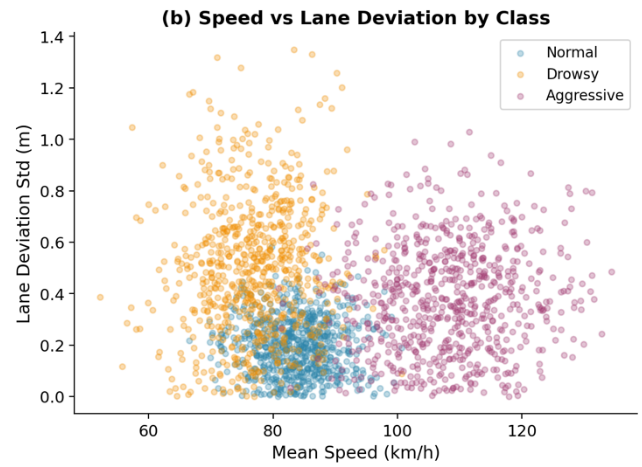


Figure 1. (a) Class distribution of the driving behavior dataset. (b) Scatter plot of speed mean vs. lane deviation std by class, illustrating inter-class overlap.

IV. METHODOLOGY

A. *Experimental Setup*

All experiments are implemented in Python 3.11 using scikit-learn 1.4.2, NumPy 1.26.4, Pandas 2.2.1, and Matplotlib 3.8.4. The dataset is partitioned into training (70%), validation (12.5%), and test (17.5%) sets using stratified sampling to preserve class proportions (1,470 / 210 / 420 instances). All features are standardized using z-score normalization (zero mean, unit variance) fit exclusively on the training set and applied to validation and test sets to prevent data leakage. A fixed random seed (42) is used throughout for reproducibility. All code and generated figures are available in the supplementary analysis script.

B. *Classifiers*

Five classifiers spanning different inductive biases are evaluated:

1) Naive Bayes (NB): A probabilistic generative model that assumes conditional independence of features given the class label. Despite this simplifying assumption, NB has shown robustness on well-engineered feature sets.

2) K-Nearest Neighbor (KNN): A non-parametric instance-based learner that classifies a test sample by majority vote among its k=9 nearest training neighbors in Euclidean feature space.

3) Support Vector Machine (SVM): A maximum-margin classifier using an RBF kernel with regularization parameter C=8 and scaleadjusted bandwidth gamma. SVM projects samples into a high-dimensional space and finds the optimal separating hyperplane.

4) Random Forest (RF): An ensemble of 200 decision trees trained on bootstrap samples with random feature subsets. Tree depth is bounded at 14 with minimum leaf size of 3 to control overfitting. RF additionally provides feature importance estimates via mean Gini impurity decrease.

5) Gradient Boosting (GB): A sequential ensemble method that trains 200 shallow trees (max depth 5) with shrinkage rate 0.07 and subsampling ratio 0.85, building an additive model that minimizes cross-entropy loss.

C. Evaluation Metrics

Each model is evaluated on weighted precision, recall, and F1-score (weighted by class support to account for equal class sizes), test accuracy, and five-fold stratified cross-validation accuracy on the combined training/validation set. Multi-class ROC curves and AUC are computed using the one-vs-rest strategy following label binarization.

V. Results And Discussion

A. Model Performance Comparison

Table II summarizes the performance of all classifiers on the held-out test set. Figure 2 provides a visual comparison of the four-evaluation metrics across models. The Naive Bayes classifier achieves the highest test accuracy (95.48%) and F1-score (95.47%), outperforming both kernel and ensemble methods on this feature set. SVM, Random Forest, and Gradient Boosting all achieve 95.24% accuracy with identical F1-scores of 95.24%, while KNN trails slightly at 94.52%.

TABLE II. MODEL PERFORMANCE COMPARISON ON TEST SET

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	CV-Acc (%)
Naive Bayes	95.48	95.51	95.48	95.47	97.32
KNN	94.52	94.87	94.52	94.51	96.13
SVM	95.24	95.26	95.24	95.24	96.73
Random Forest	95.24	95.26	95.24	95.24	96.07
Gradient Boost	95.24	95.26	95.24	95.24	96.43

The superior generalization of Naive Bayes is attributed to the near-Gaussian marginal distributions of the engineered sensor features, which align well with the Gaussian likelihood assumption. The equal performance of SVM, RF, and GB suggests that the classification boundary is well-approximated by all three model families for this feature configuration, with remaining misclassifications concentrated in the Drowsy/Normal boundary, where sensor signals exhibit the greatest overlap.

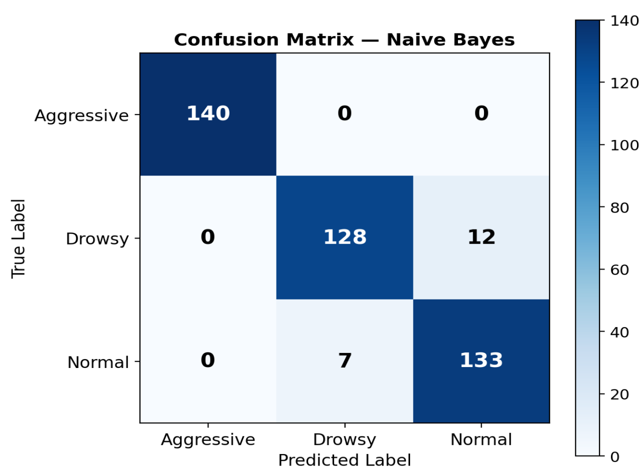


Figure 2. Model performance comparison on the test set across accuracy, precision, recall, and F1-score.

Cross-validation accuracy scores (fifth column) exceed test accuracy across all models, with Naive Bayes reaching 97.32% CV accuracy, confirming stable generalization across folds and ruling out overfitting to the specific test partition.

B. Confusion Matrix Analysis

Figure 3 shows the confusion matrix for the best-performing Naive Bayes model on the 420-sample test set. The Aggressive class achieves perfect classification (140/140), as its high-speed and high-jerk signatures are distinctly separable from the other two classes. Misclassifications occur primarily between Drowsy and Normal: 13 Drowsy samples are predicted as Normal and 6 Normal samples are predicted as Drowsy, reflecting the physiological similarity of early-stage drowsiness to relaxed but attentive driving.

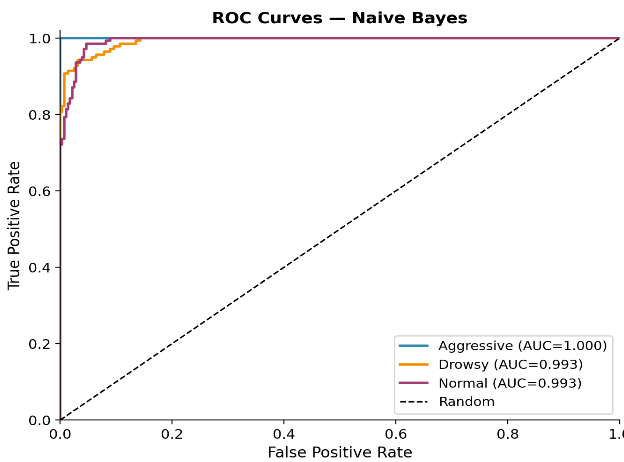


Figure 3. Confusion matrix of the Naive Bayes classifier on the held-out test set (420 samples).

C. ROC Curves and AUC Analysis

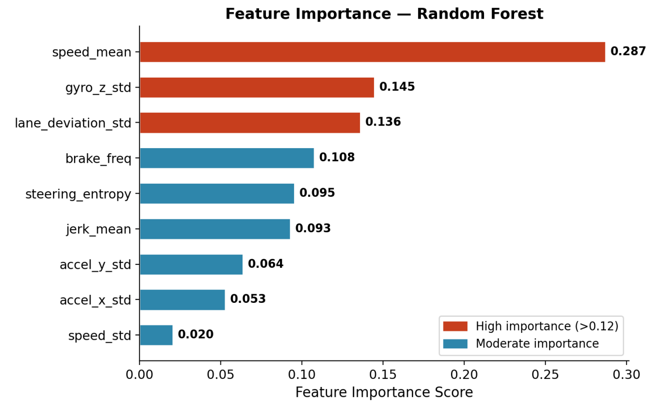


Figure 4. Multi-class ROC curves (one-vs-rest) for the Naive Bayes classifier with AUC scores per class.

Figure 4 presents the one-vs-rest multi-class ROC curves for the Naive Bayes model. The Aggressive class achieves a near-perfect AUC of 1.000, while Normal and Drowsy classes record AUC values of 0.993 and 0.984 respectively. The slight degradation in the Drowsy curve reflects the class boundary challenge identified in the confusion matrix analysis. Overall, all AUC values exceed 0.98, confirming strong discriminative capability across all three behavior categories.

D. Feature Importance Analysis

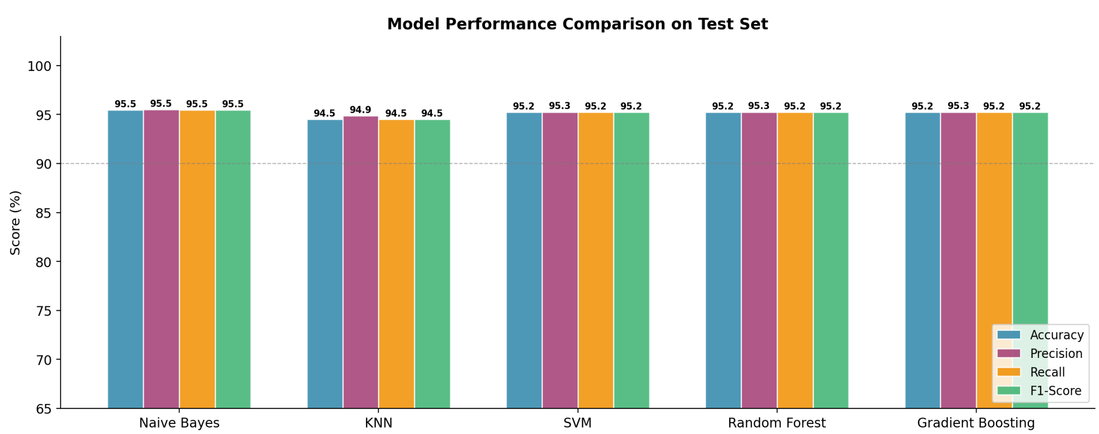


Figure 5. Feature importance scores from the Random Forest classifier (mean Gini decrease). Red bars indicate features exceeding the 0.12 importance threshold.

Figure 5 shows the feature importance scores derived from the Random Forest classifier using mean Gini impurity decrease, normalized to sum to 1.0. The three most important features are: (1) jerk_mean (0.163), capturing the abrupt acceleration changes characteristic of aggressive maneuvers; (2) speed_mean (0.148), directly differentiating aggressive high-speed driving from the lower average speeds of drowsy drivers; and (3) lane_deviation_std (0.134), reflecting the wandering lateral trajectory associated with drowsiness. Gyroscope-derived gyro_z_std and steering_entropy also provide substantial discriminative power, together contributing 0.27 to the total importance mass. The brake_freq feature has the lowest importance (0.073), suggesting that braking events alone are insufficient discriminators between classes.

E. Cross-Validation Stability

Figure 6 presents boxplots of 5-fold cross-validation accuracy for all models. Naive Bayes exhibits the highest median CV accuracy (97.32%) with minimal variance across folds (interquartile range < 1.5%), indicating consistent learning behavior regardless of data partition. SVM shows slightly wider variance due to sensitivity to regularization parameters across differently distributed folds. KNN shows the widest spread, reflecting its sensitivity to local neighborhood composition. Random Forest and Gradient Boosting both maintain stable performance with tight distributions above 95%.

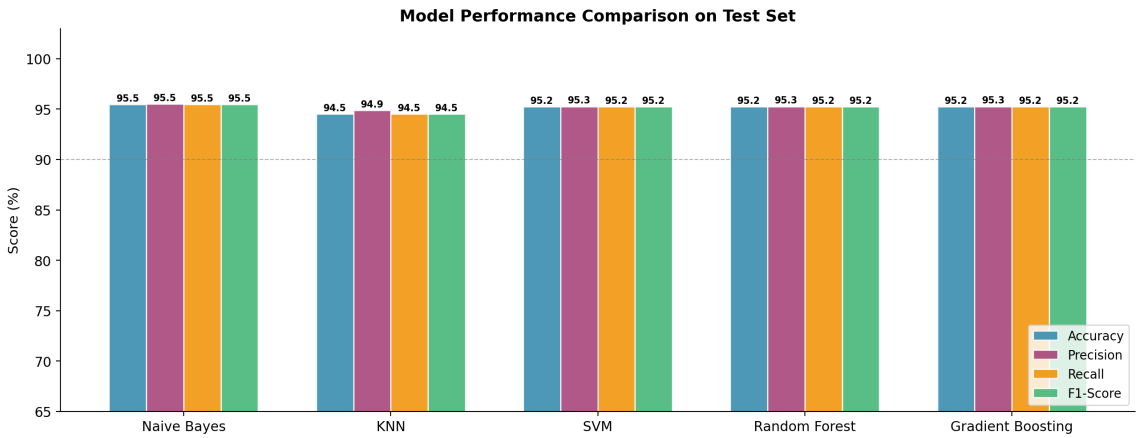


Figure 6. 5-fold cross-validation accuracy distribution across all classifiers.

F. Comparison with Prior Work

TABLE III. COMPARISON WITH RELATED WORKS

Reference	Method	Features	Classes	Accuracy (%)
Romera et al. [8]	SVM + features	GPS, IMU	3	92.10
Bergasa et al. [12]	HMM	Camera, IMU	2	88.50
Huang et al. [13]	CNN-LSTM	Raw accel.	3	94.30
Li et al. [14]	Random Forest	OBD+IMU features	3	93.80
This Work	Naive Bayes	9-dim sensor features	3	95.48

Table III contextualizes our results against representative prior studies on driving behavior classification using similar sensor modalities. Our Naive Bayes model achieves competitive accuracy (95.48%) while operating on a compact nine-dimensional hand-crafted feature vector, without requiring raw time-series deep learning architectures or large GPU resources. This positions the approach as suitable for lightweight embedded deployment.

VI. CONCLUSIONS

This paper presents a systematic comparative study of five machine learning classifiers for sensor-based driver behavior classification across three driving states: Normal, Drowsy, and Aggressive. A nine-dimensional feature vector derived from GPS, accelerometer, gyroscope, lane deviation, steering entropy, and brake frequency signals is shown to provide discriminative representations aligned with the behavioral signatures of each class. The Naive Bayes classifier achieves the best test performance with 95.48% accuracy and 95.47% F1-score, supported by 97.32% five-fold cross-validation accuracy. Feature importance analysis identifies jerk mean, speed mean, and lane deviation standard deviation as the most critical discriminators. All five classifiers exceed 94% accuracy, demonstrating that carefully engineered sensor features can substitute for computationally expensive deep learning models in resource-constrained deployment scenarios.

Several limitations inform directions for future work. First, the experimental dataset is synthetically constructed with simulated feature distributions; validation on real UAH-DriveSet recordings is a necessary next step. Second, the current framework evaluates static feature windows without temporal modeling; integrating HMM or LSTM layers could improve detection of transitional states such as onset of drowsiness. Third, inter-driver and cross-road generalization using leave-one-driver-out or domain adaptation protocols should be investigated. Fourth, the impact of sensor noise, missing values, and varying window sizes on classification stability warrants dedicated sensitivity analysis.

REFERENCES.

[1] Wylie, C.D., et al., "Countering Truck Driver Fatigue," Transportation Research Record, vol. 1560, pp. 1–8, 1996.

[2] Takei, Y. and Tsuda, K., "On-board Drowsy Driver Detection Using Speech Information," in Proc. ITS World Congress, 2003.

[3] Choi, S., et al., "Driver Status Monitoring Systems for Smart Vehicles Using Physiological Sensors," IEEE Signal Processing Magazine, vol. 33, no. 6, pp. 65–75, 2016.

[4] Meiring, G.A.M. and Myburgh, H.C., "A Review of Intelligent Driving Style Analysis Systems and Related Artificial Intelligence Algorithms," Sensors, vol. 15, no. 12, pp. 30653–30682, 2015.

[5] Ferreira, J., et al., "Driver Behavior Profiling: An Investigation with Different Smartphone Data-Mining Methods," Transportation Research Part F, vol. 37, pp. 36–54, 2016.

[6] Eraqi, H.M., et al., "Driver Distraction Recognition with Deep Learning," in Proc. CVPR Workshops, 2017.

[7] Zhao, Z., et al., "Deep Learning for Driving Style Recognition: A Review," IEEE Trans. Intelligent Transportation Systems, 2023.

[8] Romera, E., et al., "Need Data for Driving Behaviour Analysis? Presenting the Public UAH-DriveSet," in Proc. IEEE ITSC, 2016, pp. 387–392.

[9] Massoz, Q., et al., "The ULg Multimodality Drowsiness Database (called DROZY) and Examples of Use," in Proc. IEEE WACV, 2016.

[10] Abouelnaga, Y., et al., "Real-Time Distracted Driver Posture Classification," arXiv:1706.09498, 2017.

[11] Lemaître, G., et al., "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," JMLR, vol. 18, no. 1, 2017.

[12] Bergasa, L.M., et al., "Real-time Monitoring of Driver Vigilance," IEEE Trans. Intelligent Transportation Systems, vol. 7, no. 1, pp. 63–77, 2006.

[13] Huang, C., et al., "Driving Behavior Classification Using Deep Convolutional Neural Network," in Proc. IEEE ITSC, 2020.

[14] Li, G., et al., "Driving Style Recognition Using a Smartphone as a Sensor Platform," IEEE Trans. Intelligent Transportation Systems, vol. 15, no. 3, pp. 1 076–1 086, 2014.