ISSN 2579-7069

Classifying Driver Behavior Using Machine Learning: A Simple Approach to Detect Distracted and Aggressive Driving

Ika Christine Purba[1,*], Aulia Al-JIhad Safhadi[2]

[1,2]*Magister of Computer Sciences, Universitas Gadjah Mada, Indonesia*

(Received June 8, 2025; Revised October 12, 2025; Accepted January 27, 2026; Available online March 29, 2026)

Abstract

This study explores the use of Machine Learning models to classify driver behavior as either Distracted or Aggressive, using data derived from real-world driving scenarios. Two ML algorithms, Random Forest (RF) and Support Vector Machine (SVM), were applied to classify driver behavior based on key features such as brake_pressure, lane_deviation, and headway_distance. The RF model outperformed the SVM model, achieving an accuracy of 95% compared to 94% for SVM. The study demonstrates that brake_pressure and headway_distance are the most important features for detecting Aggressive driving, while lane_deviation is crucial for identifying Distracted driving. The findings suggest that RF is particularly effective in handling complex, high-dimensional data, providing accurate and reliable predictions. The results contribute to the advancement of road safety technologies by enhancing the detection of unsafe driving behaviors, which can be integrated into Advanced Driver Assistance Systems (ADAS) and autonomous vehicles. Future work should focus on expanding the dataset, integrating more diverse sensor data, and exploring more complex ML models, such as deep learning, to further improve classification accuracy and real-time performance in real-world applications.

Keywords: Driver Behavior, Machine Learning, Random Forest, Support Vector Machine, Road Safety

1. Introduction

Road safety remains a critical concern worldwide, with over 1.3 million fatalities from road traffic accidents annually. A major contributing factor to these tragic statistics is unsafe driver behavior, which significantly increases the risk of accidents. Driver behavior, including distractions and aggressive actions, plays a pivotal role in determining road safety outcomes. As a result, there is a growing emphasis on identifying and addressing these behaviors to enhance overall traffic safety.

Recent studies have highlighted the effectiveness of advanced technologies, particularly Machine Learning (ML), in monitoring and predicting driver behavior. For instance, Gupta et al. [1] demonstrate how deep learning models can detect unsafe driving behaviors such as distraction and fatigue within Intelligent Transport Systems (ITS). These systems offer promising solutions by providing real-time insights into driver performance, which can significantly contribute to reducing traffic accidents. Similarly, research by Fanai et al. [2] showcases interventions targeting dangerous driving behaviors in low- and middle-income countries, which are crucial for improving road safety management globally.

In addition to distraction and fatigue, aggressive driving is another critical behavior that contributes to accidents. Studies such as Grinerud [3] point to the importance of safety training frameworks for Heavy Goods Vehicle (HGV) drivers, emphasizing the need for specific behavioral interventions to reduce crash risks. Furthermore, the integration of eco-safe driving behaviors through ML, as discussed by Jain and Mittal [4], demonstrates how such technologies can enhance both safety and fuel efficiency by predicting and modifying risky behaviors.

ML models are rapidly gaining traction in road safety applications, offering the ability to classify driving behaviors accurately. Fu et al. [5] demonstrate the effectiveness of a distracted driving detection model that employs deep neural

***Corresponding author: Ika Christine Purba (ikachristinepurba@mail.ugm.ac.id)**
DOI: https://doi.org/10.47738/ijiis.v9i2.299

networks to assess driver performance under various states. By simulating real-world driving conditions, their model provides actionable insights into driver distractions and emphasizes the need for real-time monitoring systems. Zhao et al. [6] take this further by using multi-sensor data analysis to detect risky behaviors such as fatigue and distraction, underscoring the importance of integrating multiple data sources for more accurate predictions.

Despite these advancements, the real-world application of ML for driver behavior classification remains challenging. Traditional methods such as camera-based systems and sensor fusion face limitations in terms of accuracy and adaptability to different environmental conditions. For instance, while Driver Monitoring Systems (DMS) rely on cameras and deep learning to detect signs of distraction or fatigue, these systems are often hindered by poor lighting conditions and privacy concerns [7]. Recent innovations in multimodal sensor integration, combining technologies like LiDAR, radar, and cameras, show promise in overcoming these limitations, yet challenges remain in managing the data complexity and ensuring real-time processing for high-stakes applications [8].

In this context, the use of ML techniques, including deep learning and ensemble methods, has the potential to address the existing gaps in driver behavior detection. Ensemble methods, as highlighted by Surapunt and Wang [9], combine multiple models to enhance prediction accuracy, particularly in uncertain and dynamic driving conditions. By employing such approaches, it is possible to improve the detection of distracted and aggressive driving behaviors in real-time, fostering safer driving environments.

The primary objective of this research is to develop a robust ML model capable of classifying driver behavior into two critical categories: distracted and aggressive. Using advanced neural network architectures, including ResNet50, this model will be evaluated against a real-world driving dataset to assess its performance in various driving conditions. The research also aims to evaluate the effectiveness of the developed model in detecting and classifying behaviors accurately, providing insights into its potential integration into driver assistance systems and autonomous vehicle technologies.

In conclusion, this study seeks to contribute valuable insights into the ongoing efforts to improve road safety through the use of ML for driver behavior monitoring. By accurately detecting distracted and aggressive driving behaviors, this research aims to enhance real-time safety interventions, ultimately reducing accident rates and improving traffic safety outcomes.

2. Literature Review

2.1. Overview of Driver Behavior Classification

Driver behavior classification plays a crucial role in improving road safety, as human factors are identified as key contributors to traffic accidents. Studies have emphasized the significance of identifying aggressive, distracted, or risky driving behaviors as a primary determinant of accidents [10]. This classification enables targeted interventions that can mitigate accident risks and enhance safety for all road users [11].

Historically, driver behavior classification has evolved from simplistic observational methods to more sophisticated ML techniques that analyze driver actions in real-time. Recent advancements leverage data from various sources, including smartphone sensors and onboard vehicle logs, to monitor driving patterns [12]. These developments show the shift toward data-driven approaches for identifying unsafe driver behaviors and facilitating interventions aimed at reducing road accidents.

Despite progress, inconsistencies in modeling approaches and outcome interpretations persist across studies, suggesting the need for standardization in driver behavior classification methodologies [13]. Nonetheless, the integration of cutting-edge technologies and methods has significantly enhanced the classification accuracy of driver behavior, with some models achieving over 90% accuracy in real-time applications [14], [15].

2.2. Machine Learning Approaches in Driver Behavior Classification

ML techniques, particularly supervised learning methods, have become pivotal in classifying driver behaviors. Random Forest (RF) and Support Vector Machines (SVM) are widely used due to their robustness and suitability for high-

dimensional data, making them well-suited for driver behavior classification tasks [16]. These models excel in capturing complex patterns in driving data, helping identify both distracted and aggressive behaviors.

In addition to traditional methods, deep learning techniques such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) have demonstrated remarkable performance in capturing intricate patterns in large datasets, particularly those with time-series or image data [17]. CNNs, for instance, are particularly effective in detecting distracted driving by analyzing facial expressions and visual cues, outperforming traditional methods like SVM and RF in image-based tasks [18].

However, deep learning models often require extensive labeled datasets and substantial computational power, which can hinder their real-time application, especially in resource-constrained environments [19]. While deep learning can automatically extract features, thereby reducing the need for manual feature engineering, the challenge lies in the trade-off between model complexity and real-time processing capabilities, which is critical in practical scenarios.

2.3. Feature Engineering for Driver Behavior

Effective feature engineering is central to the successful classification of driver behavior. Key features such as speed, acceleration, brake pressure, lane deviation, and phone usage have been identified as crucial indicators for classifying behaviors like aggression or distraction [20]. Speed and acceleration offer insights into how the driver operates the vehicle, while brake pressure is particularly indicative of aggressive behaviors like sudden stops. Lane deviation is another critical feature, as it signals a driver's adherence to road guidelines, and phone usage is a direct factor in assessing distraction risks [21].

The integration of sensor data from modern vehicles, including GPS, LiDAR, and Inertial Measurement Units (IMUs), has significantly enhanced the accuracy of real-time driver behavior analysis. These sensors collect extensive data relevant to driving conditions and behaviors, enabling the creation of more reliable classification models [22]. Additionally, cloud-based telemetry systems have facilitated the processing of large-scale data, improving the ability to distinguish between different driving patterns, such as eco-driving and aggressive driving [12].

However, the widespread use of sensor data raises challenges related to privacy concerns and the processing of large volumes of data in real-time. Despite advancements, there remains a need for ongoing research to address these challenges while maintaining the reliability and security of data collection and analysis methods [19].

2.4. Real-Time Driver Behavior Monitoring

Real-time monitoring of driver behavior has become essential for enhancing road safety. Systems that assess parameters such as acceleration, speed, lane departure, and phone use in real-time enable immediate interventions, potentially preventing accidents. These systems leverage a combination of onboard sensors and IoT technologies, which allow for continuous data collection and analysis, triggering timely alerts or corrective actions [23].

Such real-time monitoring applications extend to Advanced Driver Assistance Systems (ADAS) and autonomous vehicles. In ADAS, real-time feedback can trigger safety mechanisms such as warning signals or automatic braking to prevent accidents caused by distracted or aggressive driving [24]. For autonomous vehicles, integrating real-time driver behavior monitoring is crucial for enhancing the vehicle's ability to predict and react to driver actions, improving overall safety and system responsiveness in complex driving environments [25].

Despite these advancements, the deployment of real-time monitoring systems presents challenges, such as ensuring sensor accuracy, managing large datasets, and adapting to diverse driving conditions. Ongoing research is focused on refining these systems to improve their adaptability and robustness, addressing issues like sensor calibration, data privacy, and real-time processing limitations [26].

2.5. Gaps and Future Directions in Driver Behavior Monitoring

Despite the significant advancements in driver behavior monitoring, several challenges remain, particularly regarding the use of real-world data for model training. Real-world driving conditions introduce variables such as road surface quality, traffic congestion, and weather conditions, which often complicate the modeling process. Models trained in controlled environments may struggle to generalize effectively when applied to diverse, unpredictable scenarios [27].

To enhance the robustness of driver behavior classification systems, integrating multi-modal data sources, such as camera footage, LiDAR, and telematics data, has shown promising potential. Combining visual data with telemetry can provide a richer, more nuanced understanding of driving contexts, allowing for more accurate behavior detection and prediction [28]. Furthermore, ML models that can process these complex, multi-source datasets offer deeper insights into risky driving behaviors, ultimately improving road safety [29].

Future research should focus on overcoming the limitations of current systems, including addressing data quality issues, enhancing model generalizability, and optimizing real-time processing capabilities. Furthermore, integrating these systems into autonomous driving frameworks, which can adapt to unexpected driving conditions, will be essential for advancing road safety [30].

3. Methodology

3.1. Dataset Description

The dataset used in this study consists of real-world driving data collected from a variety of onboard vehicle sensors and smartphone sensors. This data is crucial for classifying driver behaviors, specifically categorizing them into two major types: Distracted and Aggressive driving. The key features extracted from the dataset provide comprehensive insights into the driver's actions and the vehicle's status, enabling accurate classification. The dataset was sourced from Kaggle, a popular platform for open datasets, and is designed to simulate real-world driving conditions. Table 1 provides a detailed description of the key features used for the classification task.

Table 1. Key Features of the Driver Behavior Dataset

Feature	Description
speed_kmph	Vehicle speed in kilometers per hour. This feature provides insights into the driver's velocity, which is a critical indicator of aggressive or distracted driving.
accel_x	Acceleration along the longitudinal (x-axis) direction. This feature indicates how rapidly the driver accelerates or decelerates the vehicle, which is important for identifying aggressive driving behaviors.
accel_y	Acceleration along the lateral (y-axis) direction. This measures how much the driver turns or changes lanes, potentially indicating erratic driving patterns.
brake_pressure	The amount of pressure applied to the vehicle's brakes. Higher values suggest aggressive braking, which is often associated with aggressive driving.
steering_angle	The angle of the steering wheel, which helps assess how much the driver is turning the vehicle. Extreme or erratic steering could indicate distracted or aggressive driving.
throttle	The position of the throttle (accelerator). High throttle values reflect a driver's intention to speed up quickly, which could suggest aggressive driving.
lane_deviation	The deviation from the center of the lane. A significant deviation can indicate distracted driving, as the driver may not be paying attention to lane discipline.
phone_usage	Whether the driver is using a mobile phone (1 = Yes, 0 = No). This feature is essential for detecting distracted driving, as phone use is a common cause of inattention on the road.
headway_distance	The distance between the vehicle and the car ahead. Short headway distances may indicate tailgating or aggressive driving, whereas larger gaps suggest safer driving practices.
reaction_time	The time it takes for the driver to react to an event or stimulus. Longer reaction times can be a sign of distraction or fatigue, both of which compromise driving safety.
behavior_label	The target variable representing the driver's behavior: **Distracted** or **Aggressive**. This is the output variable that the ML model aims to predict based on the input features.

These features were carefully selected to provide a holistic view of driving behavior. The dataset includes continuous numerical variables (e.g., speed, acceleration, brake pressure) and categorical variables (e.g., phone usage, behavior label), both of which contribute to the classification task.

3.2. Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for ML. Initially, missing values were addressed by imputing numerical features with their respective means, and rows with missing target labels (behavior_label) were removed to ensure the dataset remained complete. For feature engineering, categorical variables like behavior_label were encoded into numerical values (0 for Distracted and 1 for Aggressive), and new features such as reaction_time were created to offer more insights into driver behavior. To ensure all features contribute equally to the model, the

 ISSN 2579-7069

dataset was normalized using StandardScaler, which standardizes the features to have a mean of 0 and a standard deviation of 1. Feature selection was then performed, where key features like speed_kmph, lane_deviation, and phone_usage were identified as significant indicators of driving behavior, while redundant features were removed. RF was used to assess the importance of each feature, retaining the most relevant ones for training. Finally, the dataset was split into 80% for training and 20% for testing, using a stratified split to maintain the class distribution of Distracted and Aggressive behaviors. These preprocessing steps ensured the dataset was clean, balanced, and ready for ML model training.

3.3. Machine Learning Algorithms for Cancer Classification

In the classification of driver behavior, several ML algorithms, particularly RF and SVM, are commonly utilized due to their effectiveness and adaptability to the complexities of driving data.

RF is particularly advantageous because it can efficiently handle high-dimensional datasets with mixed features and is robust against overfitting, making it suitable for the varying patterns seen in driver behaviors [31]. Its ensemble approach enhances prediction accuracy by aggregating the outputs of multiple decision trees, leading to a comprehensive evaluation of data attributes. Each decision tree in a RF is trained on a random subset of the data and uses a random subset of features, and the final prediction is determined by averaging the results (for regression) or through majority voting (for classification). The RF model can be mathematically represented as:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t\,(x) \tag{1}$$

$\hat{y}$ is the predicted output (classification label), T is the total number of decision trees in the forest, $h_t(x)$is the prediction from the t-th decision tree.

The ability to aggregate predictions from multiple trees allows RF to achieve high accuracy and robustness, even in the presence of noisy or unbalanced data, making it ideal for driver behavior classification.

SVM on the other hand, is valuable for its ability to delineate complex decision boundaries in high-dimensional spaces. This is crucial in driver behavior classification, where the interactions between drivers, vehicles, and the environment can be intricate and non-linear [32]. SVM is particularly effective when the data is not excessively large, and it performs well with a relatively straightforward implementation, making it accessible for preliminary modeling tasks [33]. SVM works by finding a hyperplane that best separates data into different classes, maximizing the margin between the classes, which is represented by the following optimization problem:

$$\min_{\mathrm{w},b} \frac{1}{2} \|\, \mathrm{w} \,\|^2 \tag{2}$$

subject to:

$$y_i(\mathrm{w}^T \mathrm{x}_i + b) \geq 1, i = 1,2,\ldots,n \tag{3}$$

w is the weight vector that defines the hyperplane, b is the bias term, $\mathbf{x}_i$ are the input data points, y_i is the label of the i-th data point ($y_i \in \{-1,1\}$), n is the number of training samples.

The goal is to maximize the margin between the two classes while minimizing classification error. This formula illustrates how SVM aims to find the optimal hyperplane by minimizing the magnitude of the weight vector w while ensuring that each data point is correctly classified.

The choice of these models is justified by several factors: their relative simplicity in implementation, effectiveness in classifying complex, high-dimensional data, and their capacity to generalize well to unseen instances. This is particularly relevant in driver behavior datasets, which often encompass a variety of features like speed, acceleration, and more nuanced indicators of behavior [33]. Thus, employing both RF and SVM creates a balanced framework that can capture the subtleties involved in driving patterns, ultimately enhancing predictive accuracy and safety outcomes [34].

 ISSN 2579-7069

3.4. Model Evaluation for Driver Behavior Classification

Effective evaluation metrics are essential for assessing the performance of ML models in driver behavior classification. These metrics provide insights into the model's accuracy and its ability to generalize across unseen data. Key evaluation metrics used for model performance include accuracy, precision, recall, and F1-score, each providing a distinct measure of the model's ability to classify behavior correctly.

Accuracy measures the proportion of correctly predicted instances out of the total instances. It is calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

Where:

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

While accuracy gives a general sense of model performance, it can be misleading in cases of imbalanced datasets, where one class significantly outnumbers another [35].

Precision quantifies the number of true positives divided by the sum of true positives and false positives. It indicates how many of the predicted positive instances were actually positive. High precision is crucial when the cost of false positives is high, such as in safety-critical applications like driver behavior classification. Precision is given by:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

A higher precision value signifies a lower rate of false positives, which is desirable when we want to be certain that a positive prediction (e.g., aggressive behavior) is truly correct [36].

Recall (or Sensitivity) measures the ratio of true positives to the sum of true positives and false negatives. This metric is vital in cases were failing to identify a dangerous behavior (such as aggressive driving) can have severe consequences. Recall is calculated as:

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

In situations where safety is a priority, recall is more critical because we want to ensure that as many dangerous behaviors as possible are detected, even at the cost of potentially increasing false positives [37].

F1-Score is the harmonic mean of precision and recall, providing a single score that balances the two metrics. The F1-Score is particularly useful when both false positives and false negatives are important, as in driver behavior classification. It is calculated as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

The F1-Score provides a comprehensive measure of model performance by balancing the trade-off between precision and recall [38].

To ensure the robustness and generalization of ML models, techniques such as cross-validation and train-test split are commonly employed. Cross-validation involves partitioning the data into multiple subsets (folds), training the model on some folds while testing it on others, and then averaging the results. This approach reduces variability due to the way data is partitioned and gives a more reliable estimate of the model's performance [39]. The formula for cross-validation can be expressed as:

 ISSN 2579-7069

$$\text{CV} = \frac{1}{k}\sum_{i=1}^{k} \text{Score}_i \quad (8)$$

Where:

k is the number of folds.

Score_i is the performance score for the i-th fold.

The train-test split method divides the dataset into disjoint training and testing sets, typically using a ratio such as 70-80% for training and 20-30% for testing. This method helps prevent overfitting, ensuring the model is not too tailored to the training data and can generalize well to unseen data [40]. The formula for the train-test split is simply dividing the dataset into two parts:

X_{\text{train}}, X_{\text{test}}, y_{\text{train}}, y_{\text{test}} = \text{train_test_split}(X, y, test_size=0.2)

By employing these evaluation metrics and robust validation techniques, we can assess the model's effectiveness more accurately in real-world applications. These methods ensure that the classifiers used for driver behavior classification operate reliably and can adapt to diverse driving scenarios, ultimately contributing to enhanced safety and predictive capabilities in intelligent transportation systems [12].

4. Results

4.1. Model Performance

To assess the performance of the ML models used for classifying driver behavior, we applied both RF and SVM algorithms to the dataset. The models were evaluated using a range of evaluation metrics, including precision, recall, F1-score, accuracy, and confusion matrix, to gain a comprehensive understanding of their ability to accurately classify Distracted and Aggressive driving behaviors.

The confusion matrix is a vital tool for understanding the performance of a classification model. It displays the TP, TN, FP, and FN for each class, providing a breakdown of how well the model predicted each category. For the RF model, the confusion matrix shows that the model correctly classified 1888 instances of Distracted behavior and 1916 instances of Aggressive behavior, while making 113 false predictions for Distracted and 83 for Aggressive behavior. For the SVM model, the confusion matrix reveals 1856 correct classifications of Distracted behavior and 1910 of Aggressive behavior, with 145 false positives for Distracted and 89 for Aggressive.

Figure 1 presents the RF confusion matrix, illustrating the number of correct and incorrect classifications for both classes:

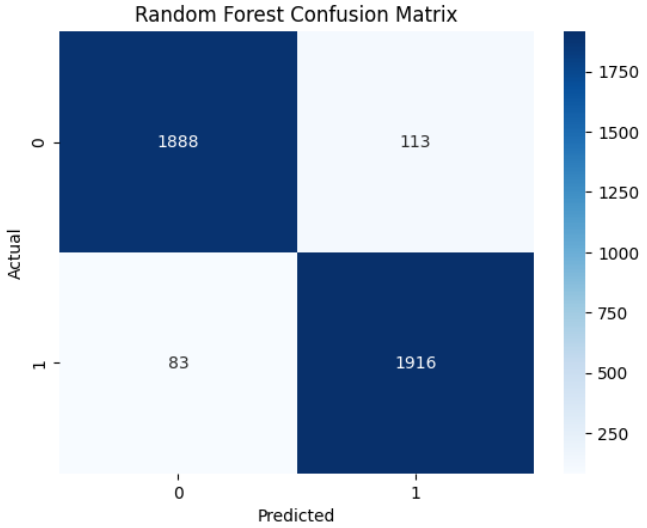

Figure 1. RF Confusion Matrix

 ISSN 2579-7069

Figure 2 presents the SVM confusion matrix, showing the distribution of true and false predictions for each behavior:

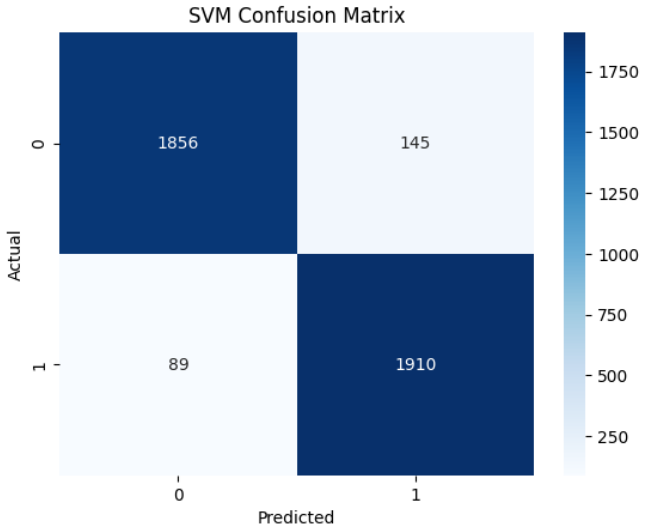

Figure 2. SVM Confusion Matrix

The classification report provides detailed metrics such as precision, recall, and F1-score for both models, offering more nuanced insights into model performance. Precision measures how many of the predicted positive instances were actually positive, while recall reflects the ability to identify all relevant instances. The F1-score is the harmonic mean of precision and recall, balancing the two metrics. Table 2 presents the classification report for the RF model.

Table 2. Classification Report for the RF Model

Metric	Distracted (0.0)	Aggressive (1.0)	Average
Precision	0.96	0.94	0.95
Recall	0.94	0.96	0.95
F1-Score	0.95	0.95	0.95
Accuracy			0.95
Macro avg	0.95	0.95	0.95
Weighted avg	0.95	0.95	0.95

Following the classification report for the RF model in Table 2, Table 3 presents the classification report for the SVM model

Table 3. Classification Report for the SVM Model

Metric	Distracted (0.0)	Aggressive (1.0)	Average
Precision	0.95	0.93	0.94
Recall	0.93	0.96	0.94
F1-Score	0.94	0.94	0.94
Accuracy			0.94
Macro avg	0.94	0.94	0.94
Weighted avg	0.94	0.94	0.94

The RF model outperforms the SVM in terms of precision, recall, and F1-score for both Distracted (0) and Aggressive (1) classes. RF achieved a higher precision (0.96 for Distracted vs. 0.95 for SVM) and recall (0.96 for Aggressive vs. 0.93 for SVM). Additionally, RF showed a higher F1-score (0.95) compared to the SVM model (0.94), indicating a more balanced performance between precision and recall.

In terms of accuracy, RF achieves an overall accuracy of 95%, slightly outperforming the SVM model, which has an accuracy of 94%. RF also excels in precision for Distracted (0) behavior (0.96 vs. 0.95 for SVM) and recall for Aggressive (1) behavior (0.96 vs. 0.93 for SVM). These differences in performance metrics demonstrate that RF is better at identifying both classes with fewer false positives and false negatives, resulting in more reliable predictions.

ISSN 2579-7069

In conclusion, while both RF and SVM models demonstrate strong performance in classifying driver behaviors as either Distracted or Aggressive, the RF model outperforms the SVM model across most evaluation metrics, including accuracy, precision, recall, and F1-score. RF is particularly effective at handling complex driver behavior classification tasks, providing more balanced predictions and better generalization across the dataset. SVM, while still a viable option, shows slightly lower performance, especially in recall for Distracted driving. Therefore, RF is recommended as the preferred model for this classification task, although SVM can still be explored for certain datasets or computational contexts.

By utilizing these evaluation metrics and comparing the models' performances, we ensure that the classifier can accurately identify distracted and aggressive driving behaviors, ultimately improving road safety through better predictive capabilities.

4.2. Feature Importance

Identifying the most important features for classification is essential in understanding the key factors that contribute to driver behavior. In this study, we used RF feature importance to assess which features are most influential in distinguishing between Distracted and Aggressive driving behaviors. RF determines feature importance by evaluating how much each feature contributes to reducing the impurity in decision trees across the ensemble.

The Feature Importance plot, shown in Figure 3, reveals that brake_pressure, lane_deviation, and headway_distance are the most important features for classifying driver behavior. Specifically, brake_pressure and headway_distance are the most prominent, with brake_pressure playing a critical role in identifying Aggressive driving behaviors, such as sudden braking and tailgating. This suggests that the model relies heavily on brake_pressure to distinguish Aggressive driving from Distracted driving.

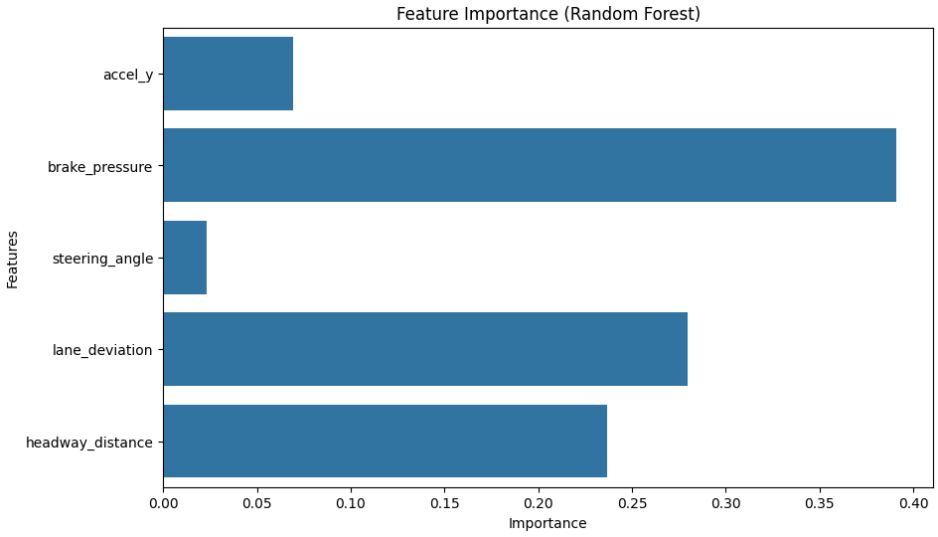

Figure 3. Feature Importance for the RF Model

Lane deviation, with an importance score of 0.23, is particularly indicative of Distracted driving. Drivers who are not paying attention may fail to stay centered in their lane, and the model uses this feature to identify such behavior. Lane deviation is essential in detecting Distracted driving, as it is often associated with drivers being less focused on their surroundings, which contrasts with the more deliberate actions involved in Aggressive driving.

Headway distance, with an importance score of 0.22, is another key feature for detecting Aggressive driving. Short headway distances indicate that a driver is tailgating or following another vehicle too closely, a hallmark of Aggressive behavior. This feature is highly valued by the model, underscoring the importance of maintaining safe following distances to prevent accidents.

Other features such as Accel_x (longitudinal acceleration) and Accel_y (lateral acceleration) play a lesser but still important role. Accel_x measures the rate of acceleration or deceleration and is more strongly associated with Aggressive driving patterns, such as rapid acceleration or harsh braking. Accel_y, on the other hand, measures lateral

movement and is less influential in this context but still contributes to detecting erratic driving behaviors, such as sharp lane changes or turns.

Finally, steering_angle also plays a role, though it is less important compared to brake_pressure or lane_deviation. The steering_angle helps to detect Aggressive driving if there are sharp or sudden steering movements, though its contribution is secondary in this model.

In conclusion, the most significant features in this classification task are brake_pressure, lane_deviation, and headway_distance, with brake_pressure and headway_distance being particularly important for identifying Aggressive driving, while lane_deviation is primarily associated with Distracted driving. These findings highlight the utility of RF in capturing critical driving patterns and distinguishing between Distracted and Aggressive behaviors in real-world scenarios, as depicted in Figure 3.

4.3. Comparison with Baseline

To evaluate the performance of the ML models, we compared them with two simple baseline models, random guessing and a heuristic-based model.

The random guessing baseline model assumes predictions are made randomly without considering any features, leading to an expected accuracy of 50% given the balanced classes (Distracted and Aggressive) in the dataset.

The heuristic-based model predicts the majority class (e.g., Distracted) for all instances. In this balanced dataset, this model achieves an accuracy of 50-60%, depending on which class dominates.

When compared to these baselines, both RF and SVM models significantly outperform the baseline models. The RF model achieved 95% accuracy, while the SVM model achieved 94% accuracy. These models demonstrated much higher precision and recall, with RF showing better overall performance.

The improvement is evident, as ML models like RF and SVM can learn from complex patterns in the data, using features such as brake_pressure, lane_deviation, and headway_distance, whereas baseline models make predictions without leveraging such detailed information.

In conclusion, RF and SVM offer significant improvements in classification performance compared to simple baseline models, demonstrating their effectiveness in handling complex driver behavior tasks and enhancing road safety monitoring systems.

5. Discussion

The results of the RF and SVM models in classifying Distracted and Aggressive driving behaviors provide valuable insights into the effectiveness of ML in predicting driver behavior. RF consistently outperformed SVM in terms of precision, recall, and F1-score, demonstrating its ability to handle complex and high-dimensional data. Key features such as brake_pressure, lane_deviation, and headway_distance were found to be critical in classifying the behaviors accurately. Brake_pressure emerged as the most important feature for detecting Aggressive driving, aligning with previous studies that highlight aggressive behaviors such as sudden braking or tailgating [20]. Similarly, lane_deviation was a strong indicator of Distracted driving, supporting findings from other research suggesting that distracted drivers tend to drift out of their lanes [21]. Additionally, headway_distance was vital in detecting Aggressive driving, as it correlates with tailgating, a common sign of aggressive driving [41].

These findings emphasize the significance of certain driving patterns and vehicle dynamics, such as brake_pressure and lane_deviation, in predicting driver behavior. Moreover, the results highlight the utility of RF in handling real-world data that requires capturing subtle behavioral patterns. The RF model's high performance in this study is consistent with previous literature, such as the work by Mosleh et al. [14], which demonstrated over 90% accuracy with ML models for driver behavior classification, reinforcing the effectiveness of RF in such tasks. However, compared to deep learning models like CNNs, which have shown superiority in image-based tasks [12], RF offers a balanced approach that combines both complexity and interpretability, making it more suitable for real-world applications like road safety monitoring.

While the results are promising, there are several limitations that must be acknowledged. One major limitation is the relatively small dataset, with only 4000 instances of driver behavior. This limits the generalizability of the model, especially in diverse driving conditions. The quality of data is another concern, as sensor data can be noisy and may contain inaccuracies in real-world environments. Furthermore, the dataset did not include critical factors such as environmental conditions (e.g., weather, traffic density) or driver-specific factors (e.g., fatigue, experience), which are known to influence driving behavior [42]. Including these additional features could enhance the model's accuracy and provide a more comprehensive understanding of driving behaviors. Moreover, while the models demonstrated strong performance in the evaluation phase, their real-time deployment in ADAS or autonomous vehicles presents challenges. These systems require rapid processing of large volumes of data, which may be difficult with current models due to computational requirements and sensor accuracy.

Looking to the future, several improvements can be made. Incorporating additional features, such as environmental data (e.g., weather conditions, traffic density) and driver-specific factors (e.g., fatigue, age), could significantly improve the model's performance [20]. Furthermore, the integration of multi-modal sensor data, combining camera footage with telemetry data from GPS and Inertial Measurement Units (IMUs), would offer richer insights into driver behavior and allow for more accurate classification, particularly in complex driving scenarios [12]. The use of more advanced models, such as deep learning techniques (e.g., LSTMs or CNNs), may further improve classification accuracy, particularly for detecting subtle behaviors like distraction based on visual cues.

For real-time deployment, future research should focus on optimizing computational resources and improving real-time processing capabilities. This is crucial for enabling the models to operate effectively in autonomous vehicles or ADAS, where real-time decision-making is essential for safety [43]. Ensuring sensor accuracy and addressing data privacy concerns will also be key factors in making these systems reliable and scalable for real-world applications.

In summary, while this study has successfully demonstrated the potential of RF and SVM models in classifying Distracted and Aggressive driving behaviors, future work should focus on expanding the dataset, incorporating more diverse features, and exploring more complex models to improve accuracy. Additionally, ensuring the real-time deployment of these models in practical driving environments will be crucial in advancing road safety technologies and enabling their application in autonomous vehicles and driver assistance systems.

6. Conclusion

This study demonstrated the effectiveness of RF and SVM models in accurately classifying Distracted and Aggressive driving behaviors. The RF model outperformed the SVM model in key performance metrics such as precision, recall, and F1-score, making it a more suitable choice for this classification task. The analysis identified key features, including brake_pressure, lane_deviation, and headway_distance, as crucial for distinguishing between the two driving behaviors. Notably, brake_pressure played a significant role in identifying Aggressive driving, while lane_deviation was a strong indicator of Distracted driving.

The findings of this research contribute significantly to enhancing road safety through the application of ML techniques in driver behavior classification. By accurately identifying Distracted and Aggressive driving behaviors, the models can improve ADAS and autonomous vehicle safety. Real-time predictions based on these models could lead to interventions that prevent accidents, making transportation systems safer for all road users.

Looking forward, further research should focus on expanding the dataset to include a broader range of driving scenarios and additional features, such as environmental data and driver-specific variables like fatigue and experience. Exploring more complex ML models, including deep learning techniques, could also enhance classification performance. Moreover, integrating multi-modal sensor data from technologies like cameras, LiDAR, and telemetry systems will provide richer insights into driver behavior, supporting real-time deployment in autonomous vehicles and driver assistance systems.

7. Declarations

6.1. Author Contributions

Author Contributions: Conceptualization I.C.P. and A.A.S.; Methodology, I.C.P. and A.A.S.; Software, I.C.P.; Validation, I.C.P.; Formal Analysis, I.C.P.; Investigation, A.A.S.; Resources, I.C.P.; Data Curation, A.A.S.; Writing Original Draft Preparation, I.C.P.; Writing Review and Editing, I.C.P. and A.A.S.; Visualization, A.A.S. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] B. B. Gupta, A. Gaurav, K. Tai Chui, and V. Arya, "Deep Learning Model for Driver Behavior Detection in Cyber-Physical System-Based Intelligent Transport Systems," *IEEE Access*, vol. 12, no. 1, pp. 62268–62278, 2024, doi: 10.1109/ACCESS.2024.3393909.

[2] S. Fanai, E. Okyere, and K. Marfoh, "Behavior Change Interventions for Dangerous Driving Behavior in Low- and Middle-Income Countries: A Systematic Review of Interventions and Outcome Measurement Instruments," *Front. Public Heal.*, vol. 13, no. 1, pp. 1–15, 2025, doi: 10.3389/fpubh.2025.1597331.

[3] K. Grinerud, "Work-Related Driving of Heavy Goods Vehicles: Factors That Influence Road Safety and the Development of a Framework for Safety Training," *Safety*, vol. 8, no. 2, pp. 43, 2022, doi: 10.3390/safety8020043.

[4] S. Mittal, P. Korde, S. Palaniraja, N. Omase, P. Guchhait, and P. Mundra, "A Systematic Review on Recent Advancement in Electric Vehicle Technologies," *Int. Res. J. Eng. Appl. Sci.*, vol. 11, no. 4, pp. 37–44, 2023, doi: 10.55083/irjeas.2023.v11i04006.

[5] B. Fu, Q. Shang, T. Sun, and S. Jia, "A Distracted Driving Detection Model Based on Driving Performance," *IEEE Access*, vol. 11, no. 1, pp. 26624–26636, 2023, doi: 10.1109/ACCESS.2023.3257238.

[6] D. Zhao, Y. Zhong, Z. Fu, J. Hou, and M. Zhao, "A Review for the Driving Behavior Recognition Methods Based on Vehicle Multisensor Information," *J. Adv. Transp.*, vol. 2022, no. 1, pp. 1–16, 2022, doi: 10.1155/2022/7287511.

[7] M. M. Faisal, M. Varshini, and K. Muvedha, "YOLO-CNN Powered Real-Time Detection of Visual Distractions and Drowsiness in Drivers," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 4, pp. 94–101, 2025, doi: 10.32628/CSEIT2511150.

[8] R. Saber, S. Ghoniemy, and M. Al-Qutt, "Driver Behavior Detection in Time Series Decade review," *Int. J. Intell. Comput. Inf. Sci.*, vol. 23, no. 3, pp. 114–140, 2023, doi: 10.21608/ijicis.2023.192999.1254.

[9] T. Surapunt and S. Wang, "Ensemble Modeling with a Bayesian Maximal Information Coefficient-Based Model of Bayesian Predictions on Uncertainty Data," *Information*, vol. 15, no. 4, pp. 228, 2024, doi: 10.3390/info15040228.

[10] S. Yaqoob, G. Morabito, S. Cafiso, G. Pappalardo, and A. Ullah, "AI-Driven Driver Behavior Assessment Through Vehicle and Health Monitoring for Safe Driving A Survey," *IEEE Access*, vol. 12, no. 1, pp. 48044–48056, 2024, doi: 10.1109/ACCESS.2024.3383775.

[11] M. Shariful Islam, M. Abu Tareq Rony, M. Safran, S. Alfarhood, and D. Che, "Elevating Driver Behavior Understanding with RKnD: A Novel Probabilistic Feature Engineering Approach," *IEEE Access*, vol. 12, no. 1, pp. 65780–65798, 2024, doi: 10.1109/ACCESS.2024.3397725.

[12] D. I. Tselentis and E. Papadimitriou, "Driver Profile and Driving Pattern Recognition for Road Safety Assessment: Main Challenges and Future Directions," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, no. 1, pp. 83–100, 2023, doi: 10.1109/OJITS.2023.3237177.

[13] X. Qiao, X. Li, W. Ma, and Y. Lu, "Longitudinal Motion Control Algorithm for Autonomous Vehicles Taking Decisions Based on the Preceding Vehicle Behavior Pattern," *Proc. Inst. Mech. Eng. Part D J. Automob. Eng.*, vol. 239, no. 12, pp. 5316–5335, 2025, doi: 10.1177/09544070251328792.

[14] B. Mosleh, J. Hamdan, B. H. Sababha, and Y. A. Alqudah, "Embedded machine learning-based road conditions and driving behavior monitoring," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 3, pp. 2571, 2024, doi: 10.11591/ijece.v14i3.pp2571-2582.

[15] X. Wang, X. Tang, T. Fan, Y. Zhou, and X. Yang, "Commercial Truck Risk Assessment and Factor Analysis Based on Vehicle Trajectory and In-Vehicle Monitoring Data," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2678, no. 12, pp. 1428–1443, 2024, doi: 10.1177/03611981241252148.

[16] K. R. Reddy and A. Muralidhar, "Machine Learning-Based Road Safety Prediction Strategies for Internet of Vehicles (IoV) Enabled Vehicles: A Systematic Literature Review," *IEEE Access*, vol. 11, no. 1, pp. 112108–112122, 2023, doi: 10.1109/ACCESS.2023.3315852.

[17] M. Ganesan, S. Kandhasamy, B. Chokkalingam, and L. Mihet-Popa, "A Comprehensive Review on Deep Learning-Based Motion Planning and End-to-End Learning for Self-Driving Vehicle," *IEEE Access*, vol. 12, no. 1, pp. 66031–66067, 2024, doi: 10.1109/ACCESS.2024.3394869.

[18] A. Asperti and D. Filippini, "Deep Learning for Head Pose Estimation: A Survey," *SN Comput. Sci.*, vol. 4, no. 4, pp. 349, 2023, doi: 10.1007/s42979-023-01796-z.

[19] E. A. Aldhahri, A. A. Almazroi, M. H. Alkinani, M. Alqarni, E. A. Alghamdi, and N. Ayub, "GNN-RMNet: Leveraging Graph Neural Networks and GPS Analytics for Driver Behavior and Route Optimization in Logistics," *PLoS One*, vol. 20, no. 8, pp. e0328899, 2025, doi: 10.1371/journal.pone.0328899.

[20] S. Hsieh, A. R. Wang, A. Madison, C. Tossell, and E. de Visser, "Adaptive Driving Assistant Model (ADAM) for Advising Drivers of Autonomous Vehicles," *ACM Trans. Interact. Intell. Syst.*, vol. 12, no. 3, pp. 1–28, 2022, doi: 10.1145/3545994.

[21] J. Medarević, S. Tomažič, and J. Sodnik, "Simulation-Based Driver Scoring and Profiling System," *Heliyon*, vol. 10, no. 22, pp. e40310, 2024, doi: 10.1016/j.heliyon.2024.e40310.

[22] K. Kwakye, Y. Seong, S. Yi, and A. Aboah, "All You Need is Data: A Multimodal Approach in Understanding Driver Behavior," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 68, no. 1, pp. 1298–1304, 2024, doi: 10.1177/10711813241275942.

[23] G. Zhang, S. Wang, W. Feng, and W. Ouyang, "A Novel Major Accidents Evolution Model and Its Application in Chinese Industrial Accident," *Heliyon*, vol. 9, no. 9, pp. e19684, 2023, doi: 10.1016/j.heliyon.2023.e19684.

[24] N. Ashrafi, S. Yousefi, G. R. Aby, S. F. Issa, H. Darabi, K. Alaei, G. Placencia, and M. Pishgar, "AI-Driven Solutions to Improve Safety and Health: Application of The REDECA Framework for Agricultural Tractor Drivers," *PLOS Glob. Public Heal.,* vol. 5, no. 6, pp. e0003543, 2025, doi: 10.1371/journal.pgph.0003543.

[25] C. Shen, B. Lei, C. Lu, and F. Zhou, "Research on the Effectiveness of Online Food Safety Supervision Under the Existence of Settled Enterprises' Myopic Cognitive Bias," *Heliyon*, vol. 9, no. 1, pp. e12784, 2023, doi: 10.1016/j.heliyon.2022.e12784.

[26] W. Kao, Y. Fan, F.-R. Hsu, C.-Y. Shen, and L. Liao, "Next-Generation Swimming Pool Drowning Prevention Strategy Integrating AI and Iot Technologies," *Heliyon*, vol. 10, no. 18, pp. e35484, 2024, doi: 10.1016/j.heliyon.2024.e35484.

[27] S. Moshfeghi and J. Jang, "Pattern Mining of Older Drivers' Driving Behavior Through Telematics-Data-Driven Unsupervised Learning," *IEEE Sens. J.*, vol. 25, no. 18, pp. 34894–34912, 2025, doi: 10.1109/JSEN.2025.3592817.

[28] L. Pinals, F. Guo, K. Ahn, S. Madden, and H. Rakha, "Safe Driving Is Sustainable Driving: An Interpretable Telematics Methodology," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2679, no. 11, pp. 221–233, 2025, doi: 10.1177/03611981251346774.

[29] K. Strandberg, N. Nowdehi, and T. Olovsson, "A Systematic Literature Review on Automotive Digital Forensics: Challenges, Technical Solutions and Data Collection," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1350–1367, 2023, doi: 10.1109/TIV.2022.3188340.

[30] B. A. Manko, "Erie Insurance: Monitoring Technology in the Car Insurance Market and the Issue of Data Privacy," *J. Inf. Technol. Teach. Cases*, vol. 13, no. 2, pp. 193–198, 2023, doi: 10.1177/20438869221117571.

[31] Z. Elamrani Abou Elassad, M. Ameksa, D. Elamrani Abou Elassad, and H. Mousannif, "Efficient Fusion Decision System for Predicting Road Crash Events: A Comparative Simulator Study for Imbalance Class Handling," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2678, no. 5, pp. 789–811, 2024, doi: 10.1177/03611981231192985.

[32] V. M. Ampadu, M. T. Haq, and K. Ksaibati, "An Assessment of Machine Learning and Data Balancing Techniques for Evaluating Downgrade Truck Crash Severity Prediction in Wyoming," *J. Sustain. Dev. Transp. Logist.*, vol. 7, no. 2, pp. 6–24, 2022, doi: 10.14254/jsdtl.2022.7-2.1.

[33] A. K. Mengistu, A. E. Gedefaw, N. D. Baykemagn, A. D. Walle, T. Z. Yehuala, M. A. Alemayehu, M. A. Messelu, and B. T. Assaye, "Predicting Car Accident Severity in Northwest Ethiopia: A Machine Learning Approach Leveraging Driver, Environmental, and Road Conditions," *Sci. Rep.*, vol. 15, no. 1, pp. 21913, 2025, doi: 10.1038/s41598-025-08005-2.

[34] A. Das, M. N. Khan, and M. M. Ahmed, "Deep Learning Approach for Detecting Lane Change Maneuvers Using SHRP2 Naturalistic Driving Data," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2677, no. 1, pp. 907–928, 2023, doi: 10.1177/03611981221103229.

[35] W. Wang, L. Zhang, B. Yan, and Y. Cheng, "Development of a Surrogate Safety Measure for Evaluating Rear-End Collision Risk Perception," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2679, no. 6, pp. 69–85, 2025, doi: 10.1177/03611981241311574.

[36] S. Mukherjee, A. D. McDonald, S. Kesler, H. Cuevas, C. Swank, A. Stevens, T. K. Ferris, and V. Danesh, "Driving Among Individuals with Chronic Conditions: A Systematic Review of Applied Research Using Kinematic Driving Sensors," *J. Am. Geriatr. Soc*., vol. 72, no. 4, pp. 1242–1251, 2024, doi: 10.1111/jgs.18738.

[37] Z. Gao, M. Bao, T. Cui, F. Shi, X. Chen, W. Wen, F. Gao, and R. Zhao, "Collision Risk Assessment for Intelligent Vehicles Considering Multi-Dimensional Uncertainties," *IEEE Access*, vol. 12, no. 1, pp. 57780–57795, 2024, doi: 10.1109/ACCESS.2024.3354383.

[38] Y. Qu, Z. Li, Q. Liu, M. Pan, and Z. Zhang, "Crash/Near-Crash Analysis of Naturalistic Driving Data Using Association Rule Mining," *J. Adv. Transp*., vol. 2022, no. 1, pp. 1–19, 2022, doi: 10.1155/2022/6562649.

[39] Y. Liang, C. Chai, W. Yin, S. Weng, and Z. Yin, "Takeover Performance Prediction Based on Function-Based Causal Inference in Level-3 Autonomous Vehicles," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2680, no. 1, pp. 396–415, 2026, doi: 10.1177/03611981251356997.

[40] G. Xu, A. Saroj, C. (Ross) Wang, and Y. Shao, "Developing an Automated Microscopic Traffic Simulation Scenario Generation Tool," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2679, no. 11, pp. 650–672, 2025, doi: 10.1177/03611981251349433.

[41] S. Bouhsissin, N. Sael, and F. Benabbou, "Driver Behavior Classification: A Systematic Literature Review," *IEEE Access*, vol. 11, no. 1, pp. 14128–14153, 2023, doi: 10.1109/ACCESS.2023.3243865.

[42] J. Yarlagadda and D. S. Pawar, "Heterogeneity in the Driver Behavior: An Exploratory Study Using Real-Time Driving Data," *J. Adv. Transp*., vol. 2022, no. 1, pp. 1–17, 2022, doi: 10.1155/2022/4509071.

[43] P. Teixeira, S. Sargento, P. Rito, M. Luis, and F. Castro, "A Sensing, Communication and Computing Approach for Vulnerable Road Users Safety," *IEEE Access*, vol. 11, no. 1, pp. 4914–4930, 2023, doi: 10.1109/ACCESS.2023.3235863.