applied sciences

MDPI

Article

Improving Driving Style in Connected Vehicles via Predicting Road Surface, Traffic, and Driving Style

Yahya Kadhim Jawad [1,*] and Mircea Nitulescu [2,*]

[1] Doctoral School, University of Craiova, 200585 Craiova, Romania
[2] Department of Mechatronics and Robotics, Faculty of Automation Computers and Electronics, University of Craiova, 200585 Craiova, Romania
* Correspondence: yahya.joad@gmail.com (Y.K.J.); nitulescu@robotics.ucv.ro (M.N.)

Abstract: This paper investigates the application of ensemble learning in improving the accuracy and reliability of predictions in connected vehicle systems, focusing on driving style, road surface quality, and traffic conditions. Our study's central methodology is the voting classifier ensemble method, which integrates predictions from multiple machine learning models to improve overall predictive performance. Specifically, the ensemble method combines insights from random forest, decision tree, and K-nearest neighbors models, leveraging their individual strengths while compensating for their weaknesses. This approach resulted in high accuracy rates of 94.67% for driving style, 99.10% for road surface, and 98.80% for traffic predictions, demonstrating the robustness of the ensemble technique. Additionally, our research emphasizes the importance of model explanation ability, employing the tree interpreter tool to provide detailed insights into how different features influence predictions. This paper proposes a model based on the algorithm GLOSA for sharing data between connected vehicles and the algorithm CTCRA for sending road information to navigation application users. Based on prediction results using ensemble learning and similarity in driving styles, road surface conditions, and traffic conditions, an ensemble learning approach is used. This not only contributes to the predictions' transparency and trustworthiness but also highlights the practical implications of ensemble learning in improving real-time decision-making and vehicle safety in intelligent transportation systems. The findings underscore the significant potential of advanced ensemble methods for addressing complex challenges in vehicular data analysis.

Keywords: ensemble learning; tree interpreter; random forest; explainable artificial intelligence (AI); algorithm CTCRA; GLOSA; connected vehicles; intelligent transportation systems

Citation: Jawad, Y.K.; Nitulescu, M. Improving Driving Style in Connected Vehicles via Predicting Road Surface, Traffic, and Driving Style. *Appl. Sci.* **2024**, *14*, 3905. https://doi.org/10.3390/app14093905

Academic Editors: Abdeljalil Abbas-Turki, Yazan Mualla and Mahjoub Dridi

Received: 25 March 2024
Revised: 26 April 2024
Accepted: 29 April 2024
Published: 3 May 2024

1. Introduction

In modern transportation, the integration of sophisticated vehicular technologies is significantly improving road safety and navigation efficiency. Modern vehicles not only feature advanced electronic control units (ECUs) but also make use of extensive networks of sensors and communication protocols. Among these, vehicle-to-vehicle (V2V) [1] and vehicle-to-infrastructure (V2I) [2] communications are critical. V2V communication allows vehicles to transmit data to each other, facilitating real-time traffic and road condition updates, while V2I communication connects vehicles to road infrastructure like traffic signals and roadside sensors, enabling broader data sharing and smarter traffic management. However, leveraging these data streams for predictive analytics introduces substantial challenges, particularly in deploying machine learning models capable of handling the complex dynamics of vehicular data. The primary objective of the research is to expedite the decision-making process for connected vehicles by analyzing data via machine learning. Consequently, this approach aims to decrease the fuel consumption rate of connected vehicles running on fossil fuels and shorten the time needed to recharge electric connected vehicles. In addition, in a prior study conducted by the authors, the focus was on reducing

diesel fuel consumption for trucks upon reaching intersections by ensuring smooth traffic flow without halting at intersections [3].

Traditional machine learning models, such as decision trees [4] and K-nearest neighbors [5], have long been the cornerstone of predictive analytics in vehicular data analysis [6]. While these methods have provided valuable insights, they inherently struggle with several critical issues when applied to the complex environments of connected vehicle systems. Two prominent issues are the "black box" nature of many machine learning models and the limitations of relying on single-model predictions [7].

A significant challenge with advanced machine learning models, particularly in the realm of deep learning, is their "black box" nature. While these models are often highly effective in making predictions, they typically provide little to no insight into the decision-making process. This lack of transparency can be problematic [8], especially in applications like autonomous driving and real-time traffic management, where understanding the rationale behind a decision is as crucial as the decision itself. The inability to interpret model predictions can hinder trust among stakeholders and end-users, posing a barrier to the broader adoption of these technologies in safety-critical systems [9,10]. Single machine learning models, despite their individual strengths, often exhibit specific weaknesses when faced with the vast and varied data types found in vehicular systems. For example, a decision tree might handle categorical data well and provide easy-to-understand rules but can suffer from overfitting and lack the ability to generalize across different datasets. Similarly, K-nearest neighbors excels at capturing local regularities but can perform poorly with high-dimensional data due to the curse of dimensions.

Moreover, single models do not always adequately capture the complex interactions and nonlinear relationships present in vehicular data. This can lead to suboptimal predictions and models that are brittle under varying conditions or scenarios that deviate from those seen during training. This research harnesses the power of ensemble learning [11,12] methodologies to improve the predictive accuracy of vehicular data analysis concerning driving style, road surface quality, and traffic conditions. By integrating multiple models via a voting classifier ensemble method [13], we address the limitations of individual approaches and enhance overall predictive performance. This approach is particularly pertinent for intelligent transportation systems, where the potential for ensemble learning remains largely underexplored.

Our proposed method marks a significant shift from traditional single-model techniques to an innovative ensemble framework that synergistically combines the strengths of random forest, decision tree, and K-nearest neighbors models. This framework utilizes a voting classifier for integrating diverse predictions, thereby enhancing both the reliability and scalability of the predictive outcomes. In this paper, we contribute significantly to the evolving field of intelligent transportation by demonstrating the effectiveness of ensemble learning in enhancing the predictive capabilities of connected vehicle systems. Our work underscores the importance of robust, transparent modeling techniques and sets a foundation for future advancements in the safety and efficiency of modern transportation.

1.1. Connected Vehicles

Via advancements in wireless inter-vehicle communication, connected vehicles (CVs) have the potential to revolutionize transportation systems. CVs establish communication not only among themselves but also with external environments. Three main groups categorize this communication: vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), and vehicle-to-everything (V2X) [14]. CVs serve as the everything system for the Internet of Vehicles (IoV), a mobile system integral to the next generation of Intelligent Transportation Systems. Inter-vehicle communication, specifically vehicle-to-vehicle (V2V), enables the sharing of collected information within a vehicle with nearby vehicles. This includes data from onboard sensors, control systems, and computers. Remarkably, safety (ITS), such as collision detection, blind-spot warning, and information sharing, can be enabled without the need for additional road infrastructure. CVs operate as decentralized wireless

ad hoc networks, supported by on-board units (OBUs) embedded computers in every vehicle [15,16]. These OBUs gather vehicle data and GPS location information, enabling communication between vehicles (V2V) and between vehicles and Roadside Units (RSUs). The success of modern smart vehicles depends on the integration of three key components: sensor fusion for seamless integration of car sensors; information systems that merge high-definition mapping with precise navigation, artificial intelligence, and robust signal processing; and technologies enabling communication and collaboration among vehicles. Show Figure 1.

Figure 1. Categories of connected vehicles.

1.2. Related Works

Our proposed system utilizes in-vehicle data, aligning with various studies that have explored the integration of machine learning (ML) applications with such data. For instance, research [17,18] employed SVM and neural network algorithms to identify unsafe driving behavior using in-vehicle sensor data such as vehicle speed, engine speed, and brake pedal pressure, achieving over 90% accuracy with both classifiers. Another study [19,20] developed a model to identify dangerous driving events using random forests and recurrent neural networks, analyzing data such as acceleration and engine RPM. This model categorized driving behavior into normal, moderate, and aggressive levels. A different approach [21] involved the use of k-means clustering and support vector machines to classify drivers' styles, focusing on vehicle speed and throttle opening. Near-crash prediction has also been a focus, with one study [22] utilizing vehicle kinematics data in multiple ML algorithms, including KNN and AdaBoost, for prediction. Another study [23] proposed a model to identify driver trips using historical data, employing decision trees with an accuracy range of 75% to 100%. Road condition analysis has also been a significant area of research. Researchers proposed a model [24] that detects road potholes using vehicle accelerometers and GPS data, evaluating it using k-means clustering and random forests. Computer vision experts suggested a pothole detection technique using stereo vision cameras and deep learning [25], while a comprehensive review [26] summarized various machine learning approaches for pothole detection. Researchers have also surveyed [27] low-cost road traffic detection techniques using in-vehicle sensors, exploring the use of various sensors such as infrared, accelerometers, and radars. Additional applications of in-vehicle data with ML include vehicle theft prevention [28,29], driver drowsiness prediction based on cabin air quality [30], and traffic signal detection from vehicle speed profiles [31]. Table 1 shows related works.

The authors in [20] discussed the advancement of smart city infrastructures via the adoption of machine-to-machine (M2M) communication models. This approach improved system automation by promoting seamless integration between physical and virtual elements in urban environments. They highlighted the development of in-vehicle technologies by automobile manufacturers, aimed at improving traffic management and information dissemination. By integrating vehicles, technologies, sensing devices, and surveillance cameras, the authors propose leveraging these elements to foster intelligent transportation systems (ITS). This, however, introduced new challenges, such as data latency and throughput requirements for certain smart city applications. The proposed model aimed to enhance M2M nodes by creating an interconnected system of electronic devices, communications, and sensors, serving as a bridge between smart cities and vehicles, with the potential for future expansion to various transportation modes.

Ref. [32] explored the emergence of smart cities (SC) and urban service robots (USR), underscoring their growing significance in daily life. The SC framework relied on collecting information via a network of sensors, culminating in centralized data hubs to optimize city operations and improve citizen services.

Table 1. Summary of related works utilizing in-vehicle data with ML applications.

Study [Ref.]	Focus Area	Key Aspects
[18,33]	Unsafe Driving Behavior	SVM, Neural Networks, over 90% accuracy
[19]	Dangerous Driving Events	Random Forests, RNN, categorization of Driving Behavior
[21]	Driving Style Classification	k-Means Clustering, SVM, Focus on Speed and Throttle
[22]	Near-Crash Prediction	Vehicle Kinematics, Multiple ML Algorithms
[23]	Driver Trip Identification	Historical Data, Decision Trees, 75–100% Accuracy
[24,34]	Road Pothole Detection	Vehicle Accelerometer, GPS Data, k-Means, Random Forests
[25]	Pothole Detection (Computer Vision)	Stereo Vision Cameras, Deep Learning
[26]	Review of Pothole Detection Techniques	Summary of Various ML Approaches
[27]	Low-Cost Traffic Detection	In-Vehicle Sensors like Infrared, Accelerometers, Radars
[25,26]	Vehicle Theft Prevention	-
[31]	Driver Drowsiness Prediction	Cabin Air Quality Analysis
[20]	Traffic Signal Detection	Vehicle Speed Profiles
[35,36]	comparison purposes	linear regression (LR)
[37]	predict travel time	support vector regression (SVR), linear kernel
[38]	map complex relationships	SVR, artificial neural networks (ANN)
[39]	queuing theory	machine learning, Kalman filter
[40]	driving safety	machine learning, Prediction tool 94.34%

Table 1. *Cont.*

Study [Ref.]	Focus Area	Key Aspects
[41]	vehicle detection, speed, length	Wireless sensor, Algorithms
[42]	road surface, road type	deep neural network (DNN) 94.27%
[43]	road surface	Sensors and Algorithm > 92%

As a result of developments in computer vision and artificial intelligence, USRs have evolved from performing straightforward tasks to taking on roles like hotel concierges, museum guides, and autonomous delivery drones. These digital agents utilize SC data to perform tasks effectively while also serving as mobile sensors and actuators for the SC. The authors proposed a conceptual solution to address common urban issues such as traffic congestion and parking shortages, suggesting the use of USRs in combination with blockchain technology for intelligent parking lot management.

Prior literature has proposed a variety of machine learning-based prediction methods, such as linear regression (LR), support vector regression (SVR), artificial neural networks (ANN), random forest (RF), and deep learning models. Most studies typically only use LR models for comparison purposes. SVR and ANN have emerged as the most commonly employed models due to their robust nonlinear fitting capabilities and capacity to map complex relationships. For instance, one study successfully applied SVR for the first time to predict travel time, demonstrating its suitability for traffic data analysis [37]. Some other researchers have used support vector machines and SVR with linear kernel functions, suggested travel time prediction models based on SVM regression, and support vector machines [38].

Despite the high accuracy demonstrated by certain models, their long computation time renders them unsuitable for real-time prediction. Researchers have explored hybrid models that combine machine learning methods with other techniques like the Kalman filter and queuing theory to address this issue [39]. However, while these studies aim to enhance prediction accuracy, they fail to provide information regarding confidence levels in travel time predictions. In response to this limitation, the integration of explainable artificial intelligence (AI) has been proposed, with a focus on enhancing transparency and trustworthiness. Tools like the tree interpreter facilitate the breakdown of feature forgetting factors, and different algorithms like random forest and projection pursuit contribute to predictions, reducing the time required for decision-making in connected vehicles.

The study [40] focused on enhancing prediction performance by utilizing reduced attributes as input, resulting in a higher accuracy of 94.34%. The paper [41] provides a comprehensive examination of a wireless sensor system for traffic monitoring. It includes computationally efficient and reliable algorithms for vehicle detection, speed and length estimation, classification, and time synchronization. Field studies conducted on both highways and urban roads across various scenarios and traffic conditions yielded impressive results, including 99.98% detection accuracy, 97.11% speed estimation accuracy, and 97% accuracy in vehicle classification based on length. Jefferson Menegazzo and Aldo von Wangenheim conducted a classification of road types and surfaces using data from three deep neural network (DNN) models. They achieved an average accuracy of 94.27% for learning data and 92.70% for verifying road surface classification validity, distinguishing between asphalt, gravel, or dirt road sections [42]. In order to identify road potholes, traffic lights, and parking lots, Ibtissem, Khedher, and Faiz Sami developed an algorithm that integrated with a mobile phone, achieving an accuracy of 92% [43].

The paper [44] highlights a detailed study of a Technology Roadmap on Intelligent and Connected Vehicles (2020) with a focus on its strategic value, technical content, and characteristics. They assessed the effects of this strategy on researchers, industries, and international strategies, highlighting the vehicle's technical structure as the "three rows and two columns" architecture. The methodology for the roadmap was presented via a case study, which is a technology analysis and the possible development routes.

The study [45] was focused on traffic congestion in weaving areas by suggesting different active fine lane management methods that can be applied in scenarios with Intelligent Connected Vehicles (ICVs). The authors carried out a VISSIM simulation to study the traffic flow as well as vehicle driving behavior in the urban expressway weaving zones. The study showed that if ICVs are instructed to complete lane changes before entering the weaving area, most time can be saved during this phase, and traffic delays can be eliminated, thus leveraging lane management and widening traffic capacity.

The researchers in [46] assessed the safety impacts of a variety of connected vehicle work zone advisory systems. They performed a comparative study using microsimulation experiments to demonstrate that CV-based warning systems are more effective than dynamic message signs (DMS). The study recommended the choice of optimal DMS positioning to achieve safe and high traffic flow conditions and stressed that real testing in the real world is needed to validate these suggestions.

The manuscript [47] was concerned with anomaly detection in intelligent transportation systems (ITS) served by a Wavelet Kernel Network with Omni-Scale Convolutional (WKN-OC). This method emphasizes the processing of high-frequency signals and feature extraction to enhance the system's ability to detect data anomalies traced to a cyberattack or sensor failures in Connected Automated Vehicles (CAVs). The authors validated that the WKN-OC model works on the SPMD data set and proved that the model has a high degree of accuracy when detecting multi- and mixed anomalies.

Finally, this research process [48] examined the current state of planning and control algorithms, focusing on self-driving vehicles that operate in urban areas, particularly. The authors thoroughly examined widely discussed proposed techniques, assessing their effectiveness, assumptions, and computing requirements; a side-by-side view of different solutions enlightened on the feasibility of the approach and its shortcomings that could be a determining factor for the system design, improving the safety and efficiency of self-driving vehicles.

2. Methodology

In this paper, the proposed methodology is structured around an ensemble learning framework, which is designed to enhance predictive performance by combining multiple machine learning models. The process begins with the acquisition of a dataset, specifically vehicle sensor data, stored in CSV format. These data are preprocessed to ensure consistency and quality, which includes merging multiple data frames, identifying missing values, and performing statistical analysis to understand the distributions and characteristics of the data. As shown in Figure 2. After that, the method moves on to the exploratory data analysis (EDA) phase, where different graphs like histograms and boxplots are made to help see the underlying patterns and find any numbers that do not fit the norm. This step is crucial as it directly informs the subsequent preprocessing strategies. Following the EDA, the pre-processed data undergoes further transformation, where numerical features are standardized using a "standard scaler" to ensure that they contribute equally to the model's performance [30]. The target variables, which include attributes like driving style and road surface conditions, are encoded using a "label encoder" to convert them into a format that can be processed by machine learning algorithms. The methodology's core is the application of ensemble learning via stacking. This involves training multiple base classifiers, in this case, a random forest classifier, a decision tree classifier, and a K-nearest neighbor classifier on the training data. Predictions from each classifier (Pred RF, Pred DT, and Pred KNN) serve as input features for the final meta-classifier. The meta-classifier, implemented via a soft voting mechanism in a "voting classifier", combines the predictions of the base classifiers to make a final prediction. This ensemble approach leverages the strengths of each individual classifier and often results in improved overall performance. The method includes an evaluation step that uses accuracy metrics, confusion matrices, and ROC curves to assess the strength of the individual classifiers and the ensemble model. We then utilize the best-performing model for the final predictions. Lastly, the methodology embraces the

concept of explainable (AI) by employing techniques such as the tree interpreter to provide insights into the contribution of each feature to the predictive models. This is instrumental in understanding the decision-making process of the educational model, thereby enhancing the transparency and trustworthiness of the predictions. This all-around method includes data preprocessing, EDA, model training, evaluation, and explanation. Its goal is to provide a full framework for predictive modeling in-vehicle data analysis.

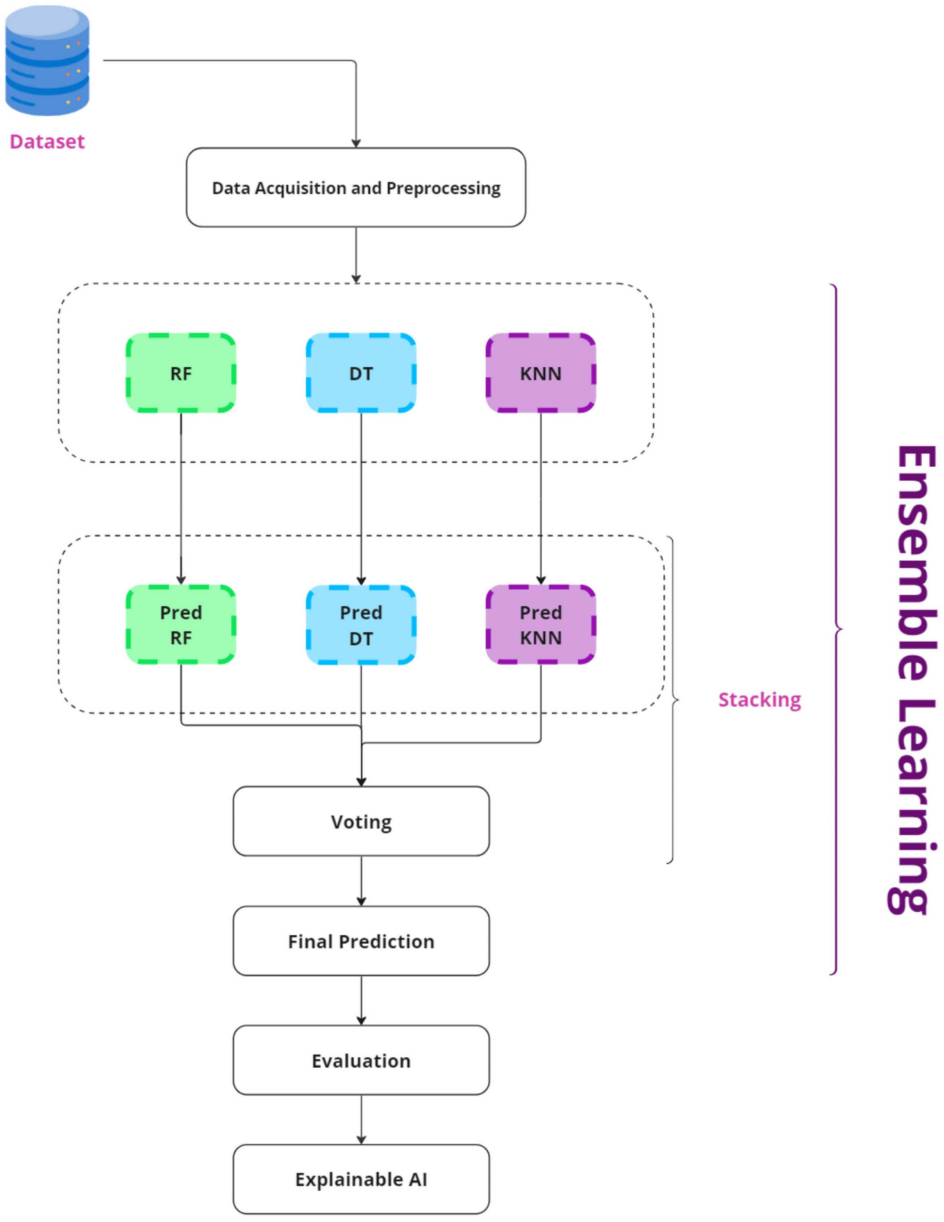

Figure 2. Proposed scheme.

2.1. Data Overview

The dataset under examination provides a detailed account of low-level parameters that are collected via the OBD-II interface from the vehicles and supplemented by data from micro-devices embedded within users' smartphones. We aim to capture a comprehensive picture of the dynamic interplay between the driver, vehicle, and environmental conditions using this rich set of data. The attributes of interest for predictive analysis encompass road surface conditions, traffic congestion levels, and driving style [31,32].

The vehicles that contribute to this dataset are the Peugeot 207 1.4 HDi and the Opel Corsa 1.3 HDi. We selected these models based on their significance in assessing the

Mafalda framework, which evaluates vehicular data. The dataset consists of fourteen numeric features that provide a multi-dimensional perspective on driving patterns. These features include an altitude change over ten seconds, instantaneous speed, and an average of speed calculations over the last minute. Speed variance and variation are also monitored second by second, along with longitudinal acceleration that is measured via the smartphone's accelerometer and refined via a low-pass filter. Additional parameters like engine load, coolant temperature, manifold air pressure (MAP), engine RPM, mass air flow (MAF) rate, and intake air temperature (IAT) provide insights into the vehicle's mechanical performance. Measurements of vertical acceleration and average fuel consumption round off the dataset. Complementing these numerical inputs are three categorical target attributes, which classify the road surface condition as smooth, full of holes, or uneven. We categorize traffic congestion as low, normal, or high and distinguish driving styles as even pace or aggressive. This multi-faceted dataset serves as a foundation for developing models that can accurately predict driving conditions and behaviors. We obtained the dataset for this work from the Kaggle website.

2.2. Exploratory Data Analysis (EDA)

The EDA is used for several reasons, with the following being the most prominent: Exploratory Data Analysis (EDA) is a fundamental step within our methodology that provides critical insights into the dataset's structure, patterns, and anomalies, which are essential for informed feature engineering and model selection [30]. During the EDA phase, we start by concatenating datasets from different vehicle models to create a unified data frame. This is immediately followed by a series of operations aimed at understanding the data's composition to obtain statistical summaries of the features:

- We pay special attention to missing values across the dataset, as their presence can significantly skew the performance of predictive models. We also identify missing values, duplicates, and inconsistencies in the dataset. This step is crucial for data cleaning and preprocessing before applying machine learning algorithms or statistical modeling. Histograms and boxplots are then meticulously generated for each numerical feature to visualize the data distribution and identify outliers. Histograms give a sense of the data's skewness and kurtosis, while boxplots help spot outliers by depicting the interquartile range. We have dealt with the three attributes (manifold air pressure, engine coolant temperature, and intake air temperature) to depict the data distribution and identify outliers in Figures 3–5. The blue line in these figures is a kernel density estimate (KDE), which provides a smooth, continuous curve representing the distribution of the dataset. This line is crucial for identifying the overall trend and shape of the data distribution beyond the discrete bars of the histogram. In Figure 3, depicting the Engine Coolant Temperature, the KDE line clearly shows a strong peak around 80 degrees, suggesting a common operational temperature range for the majority of vehicles measured. In Figure 4, which illustrates the Manifold Absolute Pressure, the KDE line peaks sharply around 100, indicating that most vehicle engines operate within a relatively narrow pressure range during normal conditions. Lastly, in Figure 5, showing the Intake Air Temperature, the KDE line highlights a dominant peak around 10 to 20 degrees, which likely reflects typical external air temperatures encountered by the vehicles. These plots are vital for recognizing features that may require normalization or transformation before being fed into machine learning models. The EDA also extends to categorical features, where count plots provide a visual representation of the distribution of categorical data.

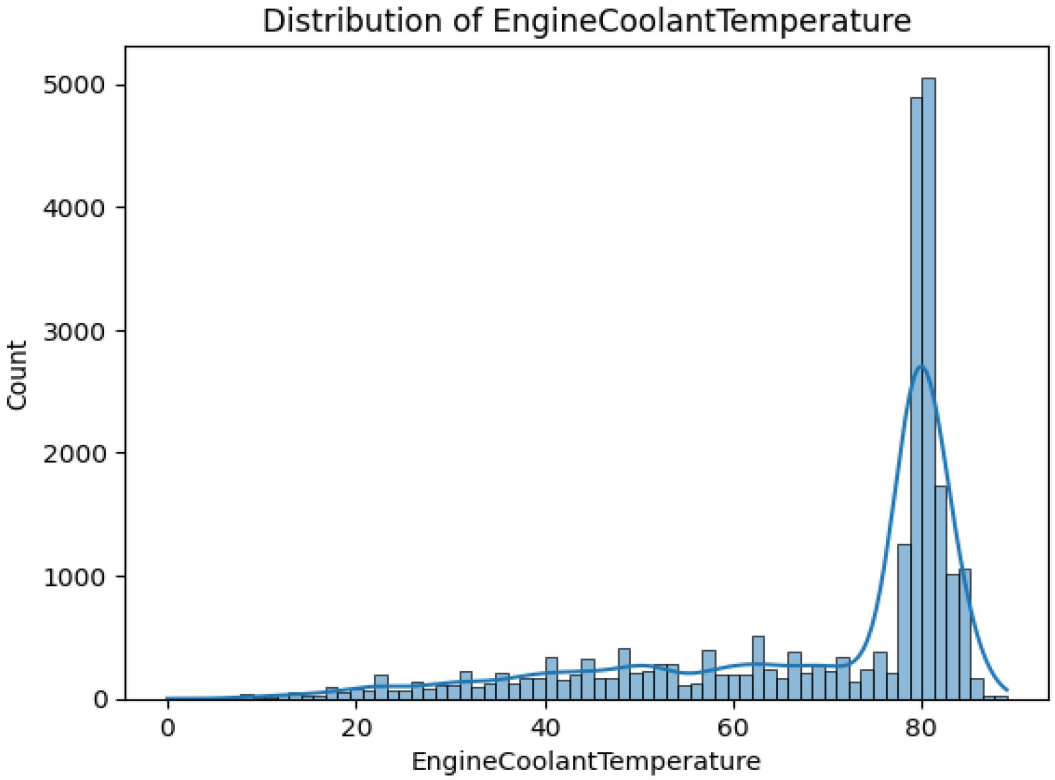

Figure 3. Distribution of engine coolant temperature.

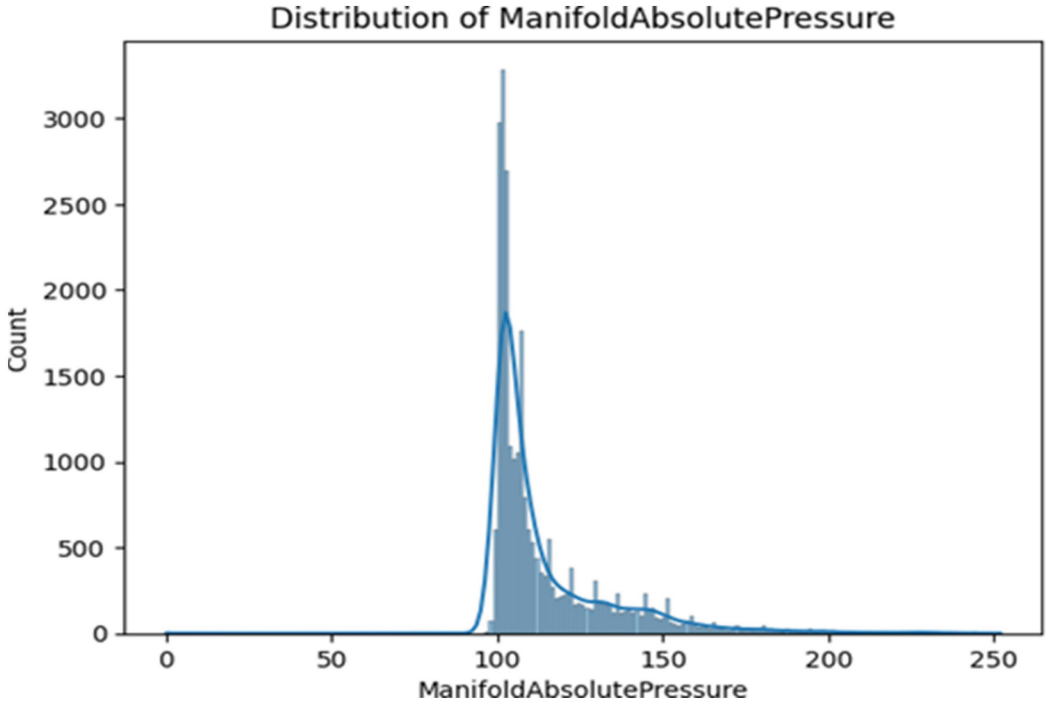

Figure 4. Distribution of manifold absolute pressure.

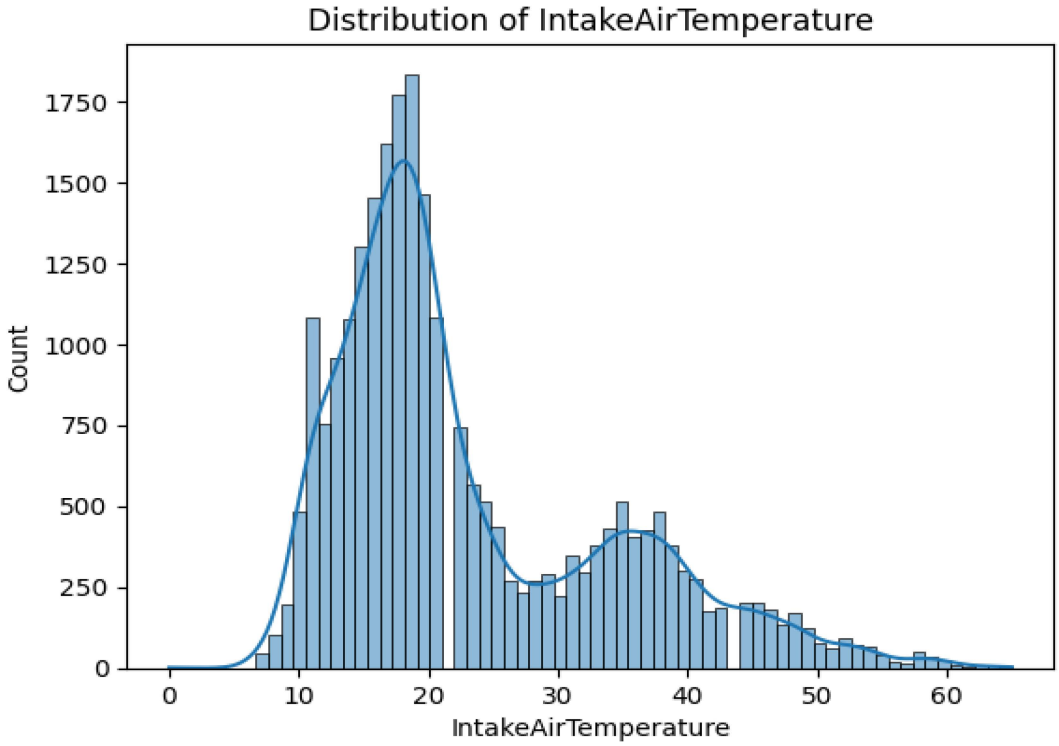

Figure 5. Distribution of intake air temperature.

- Identify relevant features or variables that have a significant impact on the target variable. It also helps in creating new features by transforming or combining existing ones to improve model performance. This aids in understanding the balance of classes within the dataset, which is important when choosing models and evaluating their performance. Furthermore, we delve into the relationships between features by employing boxplots to compare the distribution of numerical features across different categories. This not only aids in feature selection, but also in tailoring preprocessing strategies to the specific characteristics of the data.
- The EDA results serve as a foundation for communicating findings to stakeholders. Visualizations and summary statistics make complex data more understandable and facilitate decision-making processes. In essence, EDA serves as the guiding light for our preprocessing decisions, model choices, and, ultimately, the interpretability of the model outcomes. Via rigorous analysis and visual exploration, we lay the groundwork for building robust predictive models that are both accurate and understandable.

2.3. Preprocessing

In the preprocessing stage of our analysis, we undertake a series of meticulous steps to transform the raw data into a format suitable for machine learning algorithms. The initial phase involves cleaning the data by addressing any missing or inconsistent values, ensuring that the dataset is comprehensive and ready for exploration. We then proceed to normalize the numerical attributes to a common scale, eliminating any potential bias that could arise from the varying ranges of data. This normalization not only aids in improving the performance of the predictive models but also ensures that each feature contributes equally to the final predictions.

In parallel, categorical variables undergo an encoding process that converts their qualitative information into a numerical format that our algorithms can understand. This includes translating the different classes of road surfaces, traffic congestion levels, and driving styles into a structured form.

Furthermore, to prepare the data for ensemble learning, we segregate it into subsets for training and testing purposes. The training set is used to build and fine-tune the models,

while the testing set provides an unbiased evaluation of the model's performance. Via these preprocessing effects, we lay the groundwork for creating robust and effective predictive models that can accurately analyze the complex relationships within the data.

2.4. Machine Learning Models

The crux of our methodology involves the deployment of various machine learning models to predict key vehicular attributes. In our ensemble approach, we employ multiple algorithms, each bringing a unique perspective to understanding the data. We start with a random forest classifier, a model that operates by constructing a multitude of decision trees at training time and outputting the mode of the classes for classification. This model is known for its high accuracy and ability to run efficiently on large databases. Next, we utilize a decision tree classifier that creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Its simplicity and interpretability make it a valuable addition to our ensemble.

We also incorporate a K-nearest neighbors classifier, an instance-based learning method that operates on the principle of finding similar data points near each other. It's a non-parametric method, which is particularly good for capturing the complexity of the data without making assumptions about the underlying distribution. To harness the collective strength of these models, we apply a stacking approach. Here, the base classifiers are first trained on the full training set, and then their predictions are combined using a meta-classifier that makes the final prediction. Specifically, we employ a soft voting mechanism in a voting classifier.

The soft voting ensemble classifier consists of four supervised learning techniques: support vector machines (SVMs), random forests (RFs), K-nearest neighbors (KNN), and decision trees (DTs). All of these models are suitable for multi-class classification tasks. We have employed these models to compute the likelihood of categorizing the inputs as rapid progression (RP). We train each model independently and then input their respective probability predictions for progression into a logistic regression meta-classifier to produce the final output. This ensemble model is commonly known as a soft voting ensemble classifier (SVE).

This allows for a more refined decision-making process that takes into account the probability estimates from the individual classifiers instead of relying on a simple majority vote. Each individual classifier recognizes different strengths and patterns, which this ensemble method leverages to improve prediction accuracy. With each model contributing its own unique insights, the ensemble's combined predictions are typically more accurate and robust than any single model's predictions. The performance of these models is meticulously evaluated, ensuring their effectiveness and reliability in predicting driving conditions and behaviors. Via this multi-model approach, we aim to achieve a high degree of accuracy and precision in our predictive tasks, providing valuable insights into vehicular dynamics.

2.5. Ensemble Learning

Our approach to predictive modeling capitalizes on the ensemble learning technique, a process that combines the strengths of several individual machine learning models to improve the robustness and accuracy of predictions. The rationale behind this technique is that by aggregating the predictions from a group of models, we can compensate for the weaknesses of each individual model, reducing the likelihood of an erroneous prediction. In our ensemble, we utilize a collection of diverse classifiers, each with distinct characteristics.

The random forest classifier has the advantage of ensemble decision trees with bagging, which reduces overfitting while maintaining high accuracy. The K-nearest neighbors classifier helps by finding complex patterns in data without assuming how the underlying features are distributed. The decision tree classifier is easy to understand and use visually.

We then feed the predictions from these individual models into a meta-classifier that uses a soft voting strategy. This strategy considers not just the predicted classes but also

the probability of each class as predicted by each model, allowing for a more nuanced and weighted decision-making process.

By combining the models in such a manner, we take advantage of the fact that different models may make different errors, and via their combination, we can often cancel out those errors, resulting in higher overall accuracy. The final model ensemble undergoes rigorous validation using a separate testing dataset to ensure that the predictions are consistent and reliable. This ensemble learning strategy forms the cornerstone of our methodology, aiming to provide a comprehensive system that is not only accurate but also robust against diverse data scenarios.

2.6. Explainable AI

The final and critical component of our machine learning pipeline is the integration of explainable (AI) practices. Explain ability in AI is paramount, particularly in complex systems where understanding the rationale behind predictions is as important as the predictions themselves.

It is essential for establishing trust with stakeholders, refining the models, and ensuring that the decision-making process aligns with ethical standards. In the pursuit of making our machine learning models interpretable, we implement techniques that shed light on the contribution of each feature to the predictive outcomes.

By breaking down the predictions and identifying the significance of each input variable, we provide clarity on how the models arrive at their conclusions. This transparency allows us to conduct a detailed analysis of the model's behavior, detect any biases in the predictions, and provide insights that are actionable and understandable.

Moreover, explainable AI facilitates deeper collaboration between data scientists, domain experts, and end-users by making the inner workings of complex models accessible and comprehensible. It allows all stakeholders to appreciate the subtleties of the model's performance and to grasp the implications of its use in real-world scenarios. Incorporating explain ability into our methodology not only enhances the credibility of the machine learning models but also ensures that they can be audited and improved over time.

This commitment to explainable AI underscores the value we place on responsible and transparent data science practices. The tree interpreter tool, in the context of explainable AI, has been instrumental in our machine learning pipeline. This tool offers a detailed analysis of the contributions of features to specific predictions made by tree-based models, like the random forest classifier we employed in our study. In particular, we used the tree interpreter as follows:

- Feature Contribution Analysis: For each prediction, the tree interpreter divides the predicted value into contributions from each feature. This means we can see not only which features are most influential but also how much each feature sways the prediction in a particular direction, whether positive or negative.
- Stakeholder Engagement: The interpretability provided by the tree interpreter facilitates more effective communication with stakeholders. By offering clear explanations of the decision-making process, we ensured that domain experts and end-users could follow the model's logic and contribute to its improvement. Incorporating the tree interpreter into our methodology allowed for a robust cycle of analysis, interpretation, and refinement. This tool's actionable insights enabled us to improve our models' credibility and reliability. The AI's decisions, grounded in the data and aligned with the real-world phenomena it aimed to capture, provided us with assurance.

3. Results

3.1. Classification "Driving Style"

The classification results for the "driving style" attribute reveal insightful findings about the performance of various machine learning models. In our study, we focused on two primary categories of driving styles: even pace style and aggressive style.

The random forest (RF) classifier achieved a commendable accuracy of 94.67%, indicating its proficiency in identifying driving styles. The precision-recall balance was particularly impressive in classifying the predominant class, with a sensitivity of 99.05%, demonstrating the model's capability to correctly identify aggressive driving styles. The specificity remained at 59.35%, indicating moderate accuracy in recognizing even-paced driving. Show Table 2.

Table 2. Summary of classification results for "driving style" attribute.

Model	Accuracy	Sensitivity	Model	Accuracy
Random Forest (RF)	94.67%	99.05%	59.35%	0.97%
Decision Tree (DT)	91.61%	95.05%	63.88%	0.95%
K-Nearest Neighbors (KNN)	90.99%	96.24%	48.64%	0.95%
Voting Classifier	93.89%	97.93%	61.34%	0.97%

On the other hand, the decision tree (DT) classifier showed an accuracy of 91.61%. This model had a more balanced performance between the classes, with a sensitivity (recall) for the aggressive driving style at 95.05% and a specificity for the even-paced style at 63.88%. The f1-scores indicate a fair balance between precision and recall, although slightly lower than the random forest model.

The K-nearest neighbors (KNN) classifier's accuracy was slightly lower at 90.99%. It demonstrated a higher precision in predicting aggressive driving style, with a sensitivity of 96.24%. However, the even-paced driving style had a lower specificity of 48.64%, indicating some challenges in accurately identifying this class.

Lastly, the ensemble of models using a voting classifier achieved an accuracy of 93.89%. This approach, which combined the strengths of the individual models, resulted in a high sensitivity of 97.93% for the aggressive driving style, signifying a high rate of correctly identifying this class. The specificity was 61.34% for the even-paced driving style, which is an improvement over the KNN model but still suggests room for enhancement. All of these results show that ensemble methods are good at finding the right balance between sensitivity and specificity, which leads to high overall accuracy in classification tasks.

The RF model stood out with the highest accuracy and recall, indicating its strong performance in this particular scenario. The ensemble approach proved to be a robust strategy, offering a competitive alternative to individual model predictions. The two figures provided represent the feature contributions to individual sample predictions using a random forest model, as interpreted by the tree interpreter tool. Both figures visually break down how each feature influences the model's prediction of a particular class for two different samples from the dataset.

Figure 6 provides a visual depiction of the individual feature contributions to the random forest model's prediction for Sample 3. The random forest model measures and displays each feature's impact as either a positive or negative contribution towards the positive class prediction, represented by blue and red bars, respectively.

In this figure, the length of each bar signifies the magnitude of the features' influence on the model's prediction. For example, 'vehicle speed instantaneous' and 'vehicle peed average' are shown with long blue bars, indicating strong positive contributions. This suggests that for Sample 3, when these speed-related attributes have higher values, they are influential indicators that the model associates with the positive class, which could be, for instance, an 'Aggressive Driving Style'.

On the other hand, features like 'Engine RPM' and 'Mass Air Flow' exhibit red bars, denoting negative contributions. This implies that higher readings of these particular features sway the model in favor of predicting the negative class. In the context of driving style, this could mean that when the engine operates at higher RPMs, or there is greater mass air flow, the driving style may align more closely with 'Even Paced' rather than 'Aggressive' driving.

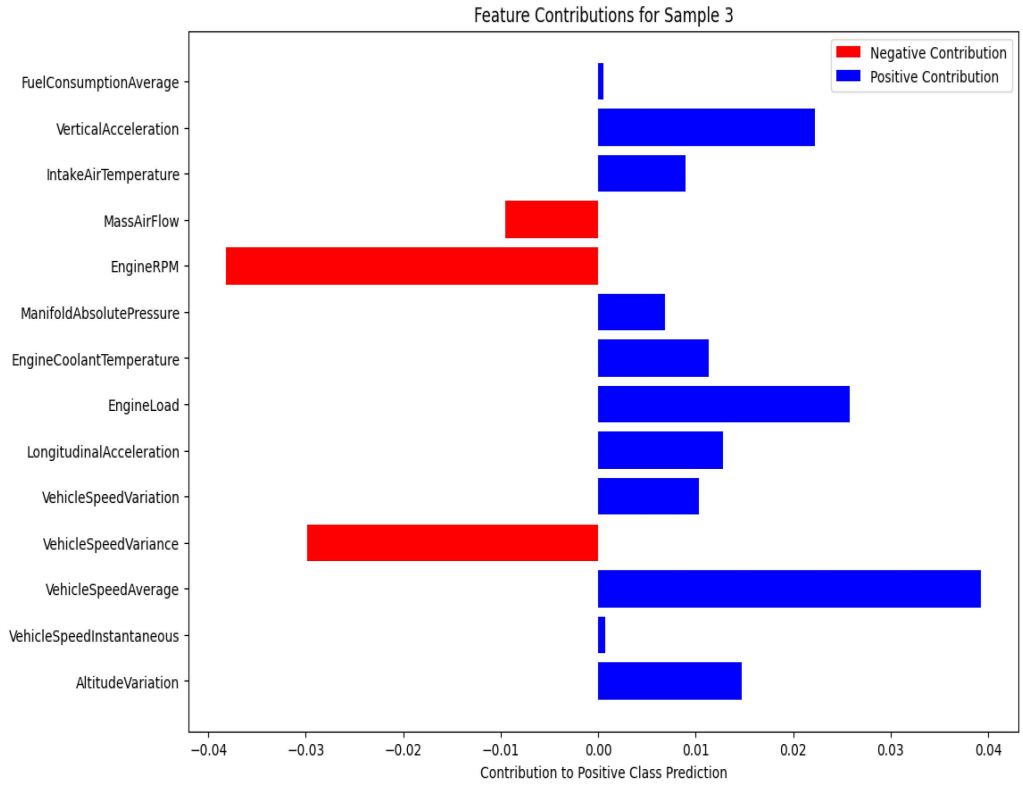

Figure 6. Feature contributions for sample 3 in driving style classification.

Understanding that these contributions are specific to Sample 3 and not universally applicable to all predictions is essential. The relative influence of these features may vary significantly across different samples, reflecting the model's complex and non-linear decision-making process. For instance, a high 'Vehicle Speed Instantaneous' may be a strong predictor for an aggressive driving style in one context, but in another, it may be less indicative due to interactions with other features or different driving conditions.

Moreover, the presence of red bars for 'Engine RPM' and 'Mass Air Flow' in this specific sample suggests a nuanced relationship between engine characteristics and driving style, highlighting the model's ability to capture the multifaceted nature of driving behavior.

Interpreting these feature contributions allows us to gain insights into the model's reasoning, fosters transparency, and builds trust in the model's predictions. However, the model's interpretability also raises the challenge of understanding the conditional nature of these contributions. Analysts must carefully consider the operational domain and the potential for feature interactions, which can introduce complexity and uncertainty in the model's predictions.

Figure 7 illustrates the impact of various features on the random forest model's prediction for Sample 20. Blue bars represent features that contribute positively to the prediction of the positive class, whereas red bars represent negative contributions that sway the prediction towards the opposite class.

In this example, features such as 'Manifold Absolute Pressure' and 'Engine Load' have prominent blue bars, indicating that they are potent indicators that strengthen the model's prediction of the positive class. This could suggest that for Sample 20, certain conditions related to engine performance and load are more likely associated with the model's identified class, which, depending on the context, might relate to specific road or driving conditions.

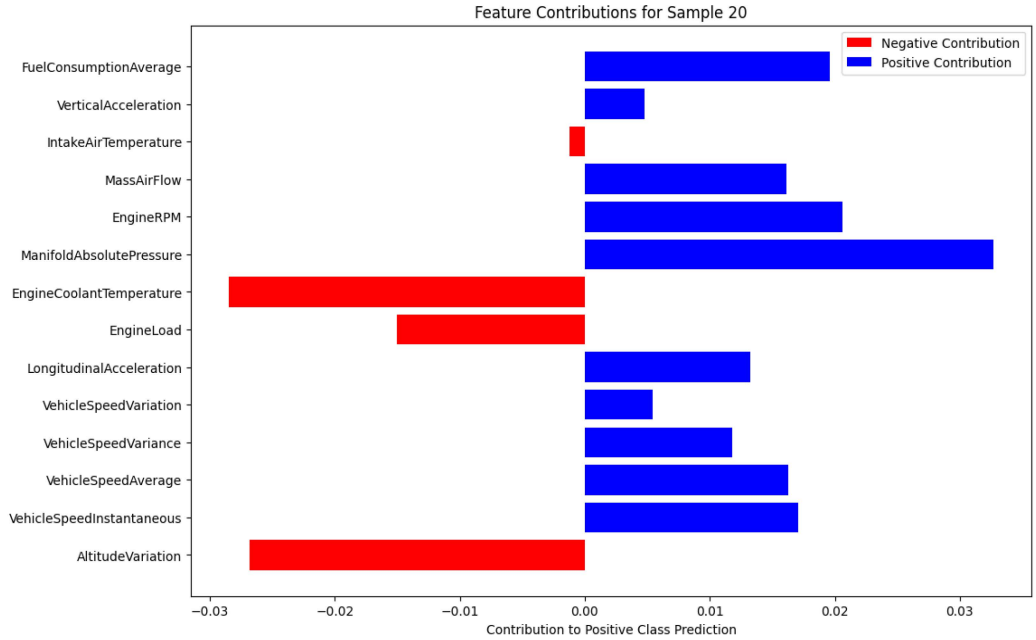

Figure 7. Feature contributions for sample 20 in driving style classification.

On the other hand, significant red bars indicate that 'Engine Coolant Temperature' and 'Engine RPM' are predictive of the negative class in this specific case. It implies that in this sample, higher engine temperatures and RPMs are less indicative of the positive class and may instead be characteristic of alternative scenarios.

The different feature contributions between Sample 3 and Sample 20 show that the model does not use a standard set of feature influences to make predictions. Instead, it changes the weight of each input based on the specifics of each sample. This adaptability underscores the complexity of the model's decision-making process and highlights the importance of considering individual sample characteristics when interpreting model predictions.

Both figures are pivotal in understanding the random forest model's decision-making process for these predictions. They offer a granular view of the model's reasoning, enabling a deeper interpretation of the model's behavior and allowing us to trace the prediction back to its influential features. This level of detail is crucial for validating the model's performance, ensuring it aligns with domain knowledge, and establishing trust in the model's outcomes by making the AI's decision-making process more transparent.

3.2. Classification "Road Surface"

The classification results for the "road surface" attribute highlight the effectiveness of the machine learning models employed in distinguishing between different road conditions. The random forest (RF) classifier emerged with an outstanding accuracy of 99.10%, demonstrating its exceptional ability to classify road surface conditions. The precision, recall, and f1-score across all conditions (full of holes condition, smooth condition, and uneven condition) were consistently high, nearly all above 99%. The sensitivity (recall) for each class was also notably high, with Full of Holes Condition at 99.25%, Smooth Condition at 99.47%, and Uneven Condition at 98.14%, indicating the model's strong capability to correctly identify each road surface condition.

The specificity values were similarly impressive, indicating the model's accuracy in ruling out the other conditions when identifying a specific one. The decision tree (DT) classifier also performed well, achieving an overall accuracy of 97.88%. While slightly lower than the RF model, the DT classifier maintained high precision, recall, and f1-scores, with macro and weighted averages ranging around 97% to 98%. The sensitivity (recall) for Full of Holes Condition was 97.30%, Smooth Condition was 98.29%, and Uneven Condition was

97.21%, all indicating high true positive rates. The specificity for these classes also remained high, reinforcing the DT classifier's reliability in road surface condition classification.

The K-nearest neighbors (KNN) classifier showed a lower accuracy of 95.35% compared to the other two models. Despite this, the KNN classifier still achieved commendable precision, recall, and f1-scores, with the macro average and weighted average scores falling between 94% and 95%. The sensitivity (recall) scores showed some variation, with Full of Holes Condition at 91.89%, Smooth Condition at 96.77%, and Uneven Condition at 93.80%, suggesting a slightly lower but still substantial ability to correctly classify the road conditions. The specificity values remained high, affirming the model's effectiveness. Finally, the ensemble model using a voting classifier achieved an accuracy of 97.26%.

The predictions made by different classifiers were combined in this model, which had high precision, recall, and f1-scores across all classes, with weighted and macro averages close to 96% to 97%. The sensitivity (recall) for the Full of Holes Condition was at 95.20%, Smooth Condition at 98.22%, and Uneven Condition at 96.05%, indicating strong positive identification rates. The specificity values were also robust, indicating solid performance in differentiating between the conditions.

In conclusion, all models demonstrated strong performance metrics, with the random forest classifier leading in almost all aspects, closely followed by the ensemble voting classifier.

These models prove to be highly effective in classifying road surface conditions, with high sensitivity and specificity across the board. Figures 8 and 9 are horizontal bar charts that illustrate the feature contributions to the random forest model's predictions for two different samples, as seen via the tree interpreter tool. Show Table 3. In Figure 8, we see the feature contributions for Sample 21 which is predicted as smooth condition class. The chart displays a range of features along the *y*-axis, with their corresponding contribution to the model's prediction on the *x*-axis. Positive contributions towards the model's prediction are shown in blue, indicating that these features push the prediction towards the smooth condition class, whereas negative contributions are in red, suggesting an influence in the opposite direction. Sample 21 is the most substantial.

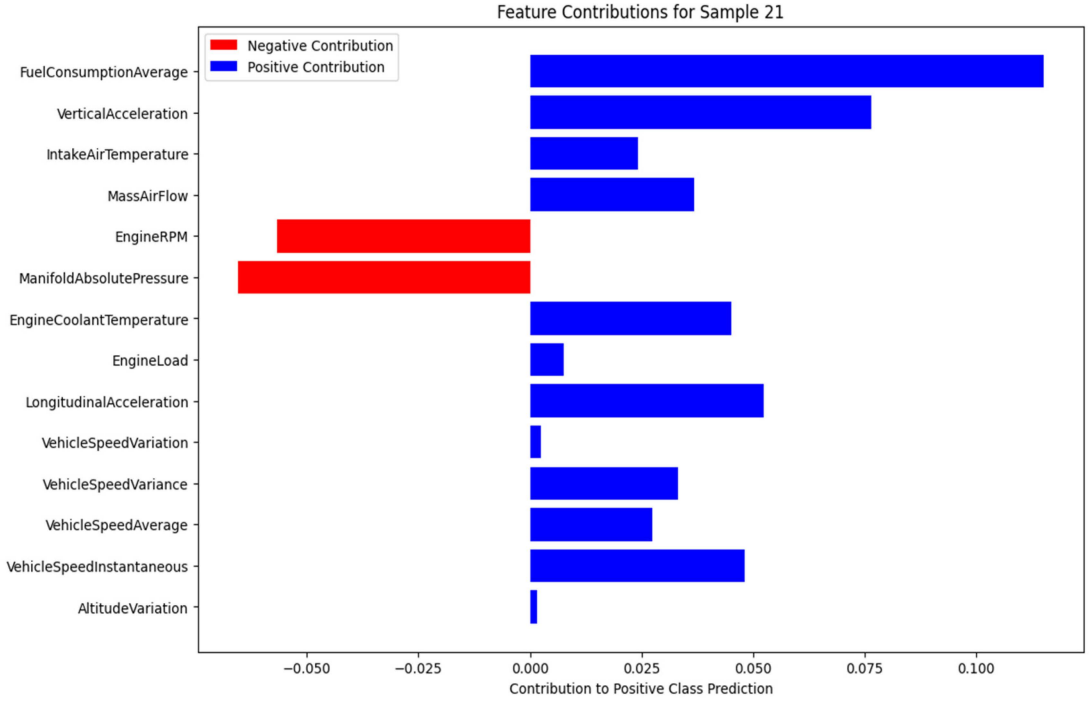

Figure 8. Feature contributions for sample 21 for road surface classification.

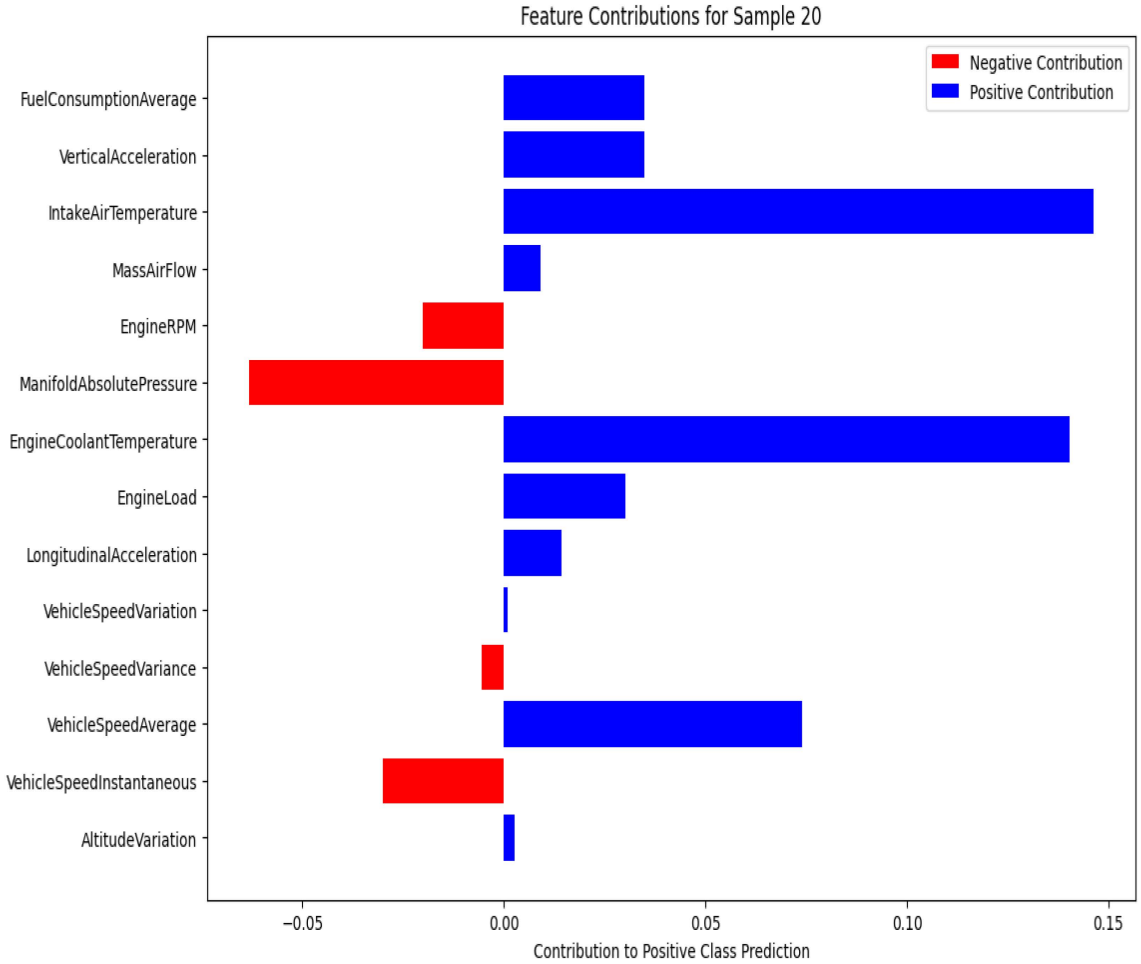

Figure 9. Feature contributions for Sample 20 for road surface classification.

Table 3. Comparison of machine learning models for "road surface" classification.

Model	Accuracy	Sensitivity (Full of Holes)	Sensitivity (Smooth)	Sensitivity (Uneven)
RF	99.10%	99.25%	99.47%	98.14%
DT	97.88%	97.30%	98.29%	97.21%
KNN	95.35%	91.89%	96.77%	93.80%
Voting Classifier	97.26%	95.20%	98.22%	96.05%

This SHAP figure reveals the factors that led the model to predict a 'smooth condition' for sample 21. high values for 'fuel consumption average', 'vertical acceleration', and 'intake air temperature' had the strongest positive influence on this classification. On the other hand, high 'engine RPM' and 'manifold absolute pressure' decreased the likelihood of the model predicting a smooth surface.

In a similar fashion, Figure 9 depicts the feature contributions for sample 20. The primary factors suggesting a 'smooth condition' are higher values for 'fuel consumption average', 'vertical acceleration', and 'intake air temperature'. Conversely, higher 'engine RPM' and 'manifold absolute pressure' values decrease the likelihood of a 'smooth condition' classification.

Both figures are critical for understanding how each feature influences the prediction of the random forest model for individual samples. They allow us to deconstruct the model's decision-making process and gain insights into which features are most predictive

for the class in question. This interpretability is crucial in validating the model's behavior and ensuring that the predictions are based on relevant and meaningful patterns in the data.

3.3. Classification "Traffic"

The classification outcomes for the "traffic" attribute illustrate the capabilities of different machine learning models to discern traffic congestion levels. The random forest (RF) classifier shows exemplary performance, boasting an accuracy of 98.80%. It exhibits high precision and recall across all traffic conditions, with the recall for high congestion conditions at 96.64%, low congestion conditions at an impressive 99.87%, and normal congestion conditions at 94.65%. The specificity for each class is also notably high, with the low congestion condition achieving almost perfect specificity at 99.95%.

The decision tree (DT) classifier, while slightly less accurate at 97.58%, still maintains strong precision and recall, with a recall of 95.64% for high congestion conditions and 98.53% for reinforcing the DT classifier's reliability in accurately identifying traffic conditions.

The K-nearest neighbors (KNN) classifier shows an accuracy of 95.75%, with the recall for the high congestion condition at 94.13%, low congestion condition at 97.49%, and normal congestion condition at 87.31%. The specificity is high for both the high congestion condition and the normal congestion condition, indicating a strong true negative rate. Show Table 4. Lastly, the ensemble model using a voting classifier achieves an accuracy of 97.14%. The model presents a high level of precision and recall, with the recall for the high congestion condition at 94.80%, the low congestion condition at 98.93%, and the normal congestion condition at 88.99%. The specificity values for the classes are also very high, especially for normal congestion conditions, at 99.38%.

Table 4. Comparison of machine learning models for "Traffic" classification.

Model	Accuracy	Recall (High)	Recall (Low)	Recall (Normal)
RF	98.80%	96.64%	99.87%	94.65%
DT	97.58%	95.64%	98.53%	93.88%
KNN	95.75%	94.13%	97.49%	87.31%
Voting Classifier	97.14%	94.80%	98.93%	88.99%

In summary, all models demonstrate high effectiveness in classifying traffic conditions. The random forest classifier stands out with the highest overall accuracy and recall, suggesting its superior performance in this classification task. The ensemble voting classifier also has strong results, which shows how important it is to combine several models to obtain high accuracy and a good balance of recall and specificity in a variety of traffic conditions. ow congestion conditions, and 93.88% for normal congestion conditions. The specificity for these classes is robust; Figures 10 and 11 depict the contribution of various features to the predictions made by a random forest classifier for two distinct samples, as analyzed by the tree interpreter tool.

In Figure 10, the horizontal bar chart shows the feature contributions for Sample 21. Each bar represents the magnitude of a feature's influence on the model's prediction, with blue bars indicating a positive contribution and red bars denoting a negative contribution. For this sample, "Vehicle Speed Instantaneous" and "Fuel Consumption Average" make the most significant positive contributions, suggesting they are highly influential in the classifier's prediction towards the positive class. Meanwhile, features such as "Engine Load" and "Manifold Absolute Pressure" exhibit the most notable negative contributions, implying they steer the prediction away from the positive class.

For Sample 20, Figure 11 provides a similar analysis. Once again, we see a combination of positive and negative contributions across different features. Notably, "vehicle speed average" shows a substantial positive contribution, while "altitude variation" and "vehicle speed variance" have strong negative impacts on the classifier's prediction. These

features have a considerable influence on the model's decision-making process for this particular sample.

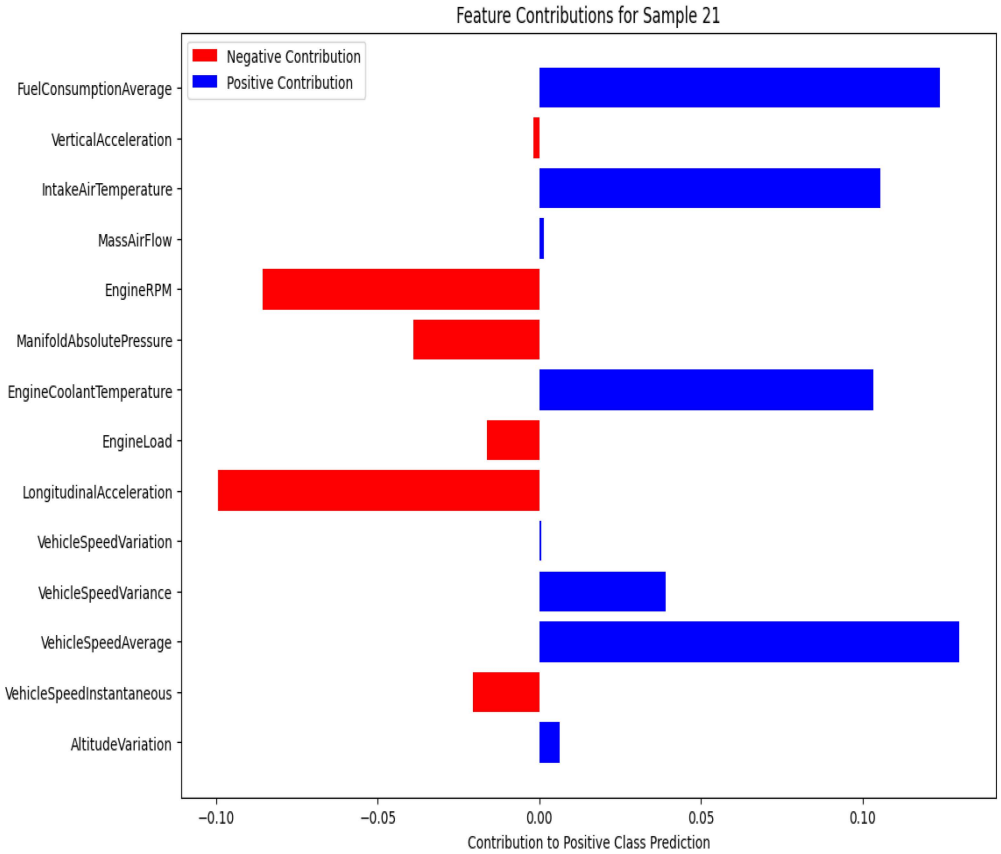

Figure 10. Feature contributions for sample 21 for traffic classification.

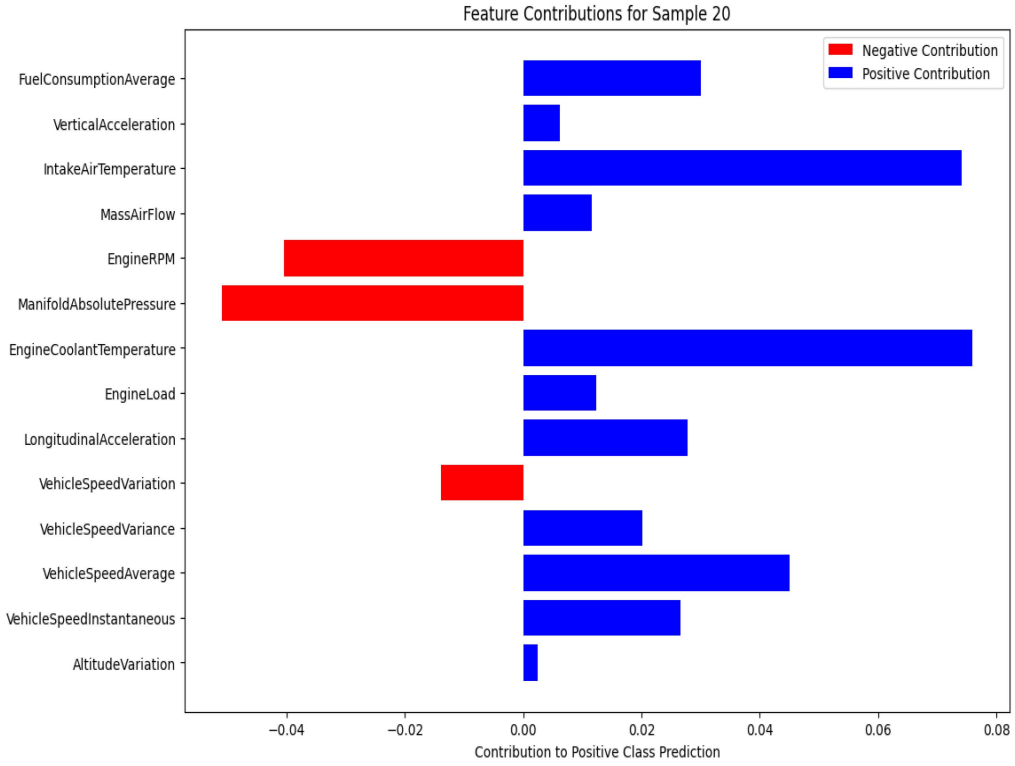

Figure 11. Feature contributions for sample 20 for traffic classification.

Both figures serve as insightful visual tools for understanding the decision-making process within the random forest model. They allow us to dissect the model's predictions and comprehend the role each feature plays in contributing to the final outcome. This understanding is crucial for model validation, offering a transparent view of the model's predictive behavior and ensuring that key features are appropriately weighted in the decision process.

From this perspective, we see the potential to leverage the findings from this manuscript. We propose integrating the features of predicting road surface, traffic, and driving style to enhance traffic congestion control along two dimensions: Firstly, integrating the results with the algorithm controlling traffic congestion in residential areas (CTCRA) to optimize routes for navigation application users. Secondly, integrating the results with the technology of Green Light Optimized Speed Advisory (GLOSA) to facilitate data sharing with connected vehicles [49], as shown in Figure 12.

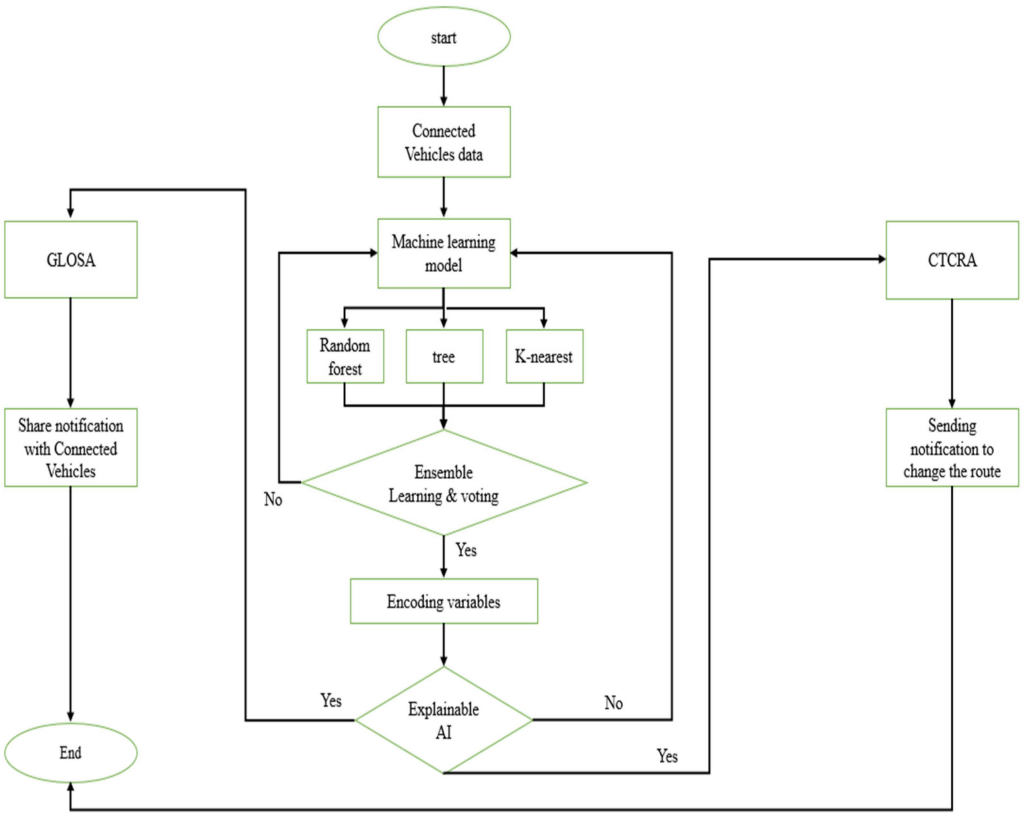

Figure 12. Algorithm improving driving style in connected vehicles (IDSCV).

3.4. Discussion

In this section, we engage in a comparative analysis of our ensemble soft voting approach against existing models in the literature across three distinct domains: driving style, road surface quality, and traffic conditions. Our ensemble method, which leverages a combination of machine learning algorithms, has shown competitive and promising results. For driving style classification, our ensemble model slightly outperformed the CNN-LSTM model referenced in [50], with an accuracy of 93.89% compared to 93.81%. This small improvement suggests that even though deep learning models like CNN-LSTM are very good, ensemble methods might be able to find more patterns by combining different algorithmic points of view. This could lead to a more accurate classification in some situations.

Using a CNN model, our work in road surface classification achieved an accuracy of 97.26%, slightly below the 97.50% reported in [51]. This small difference shows that CNNs are good at finding spatial hierarchies and patterns in da-ta, but ensemble methods are

still very useful for making more complex decisions, especially when model variance and interpretability are important factors.

With an accuracy of 97.14%, our model significantly improved traffic prediction, surpassing the 81% accuracy of the standalone random forest model in [52]. Compared to the accuracy of 98.7%, there is a slight difference [41]. This huge improvement shows how well the ensemble soft voting method works for dealing with the complexity and unpredictability of traffic conditions. Combining different models may help us understand these changing environments better.

The comparison results show that our ensemble soft voting method is very good. It not only works as well as specialized deep learning models, but it also does a much better job of predicting traffic than traditional single-model methods. It suggests a robust generalization capability and highlights the potential of ensemble methods for adapting to a variety of complex real-world conditions. Show Table 5.

Table 5. Comparison with existing works.

Ref.	Area	Model	Accuracy
[50]	Driving Style	CNN-LSTM	93.81%
Our work	Driving Style	Ensemble Soft Voting	93.89%
[51]	Road Surface	CNN	97.50%
Our work	Road Surface	Ensemble Soft Voting	97.26%
[52]	Traffic	Random Forest	81%
[41]	Traffic	Naïve Bayes	98.7%
Our work	Traffic	Ensemble Soft Voting	97.14%

4. Conclusions

Throughout this paper, we have undertaken a comprehensive exploration of ensemble learning techniques applied to vehicular data, aiming to predict driving behavior and road conditions. The experiments conducted have demonstrated the efficacy of utilizing a variety of machine learning models, such as random forest, decision tree, K-nearest neighbors, and voting classifiers, in making accurate predictions based on vehicular sensor data. Our findings reveal that the random forest classifier consistently delivers superior performance across multiple attributes, including driving style, road surface, and traffic conditions. It achieves high accuracy, sensitivity, and specificity, showcasing its robustness as a predictive tool in the automotive domain.

The ensemble voting classifier also exhibits strong predictive capabilities, reinforcing the idea that combining different models can lead to improved performance over individual classifiers. The tree interpreter tool highlights the importance of explainable AI by offering valuable insights into each feature's contribution to class prediction. This not only enhances trust in the models but also paves the way for further refinement and understanding of the driving environment. In summary, the models we have evaluated offer promising approaches for real-time vehicular data analysis, with potential applications in intelligent transportation systems, autonomous vehicle guidance, and driver assistance technologies.

The ability to accurately predict driving conditions and behavior can lead to safer, more efficient, and intelligent mobility solutions. Future work will include a comprehensive comparison with other ensemble learning techniques, such as boosting, bagging, and stacking methods. This will help to highlight the distinctive performance and characteristics of our approach in various applications. We aim to demonstrate not only the accuracy but also the computational efficiency and scalability of our method compared to these alternatives. We acknowledge the importance of interpretability in machine learning models, especially in domains where decision-making is critical.

Therefore, we plan to expand our interpretability analysis using additional case studies that cover a broader range of scenarios. We will also explore other interpretable tools and

methodologies, such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive Explanations), to provide deeper insights into our moderators' decision-making mechanisms. To address the adaptability and generalization of our model, we will undertake experiments with a more diverse dataset, encompassing a variety of road types and traffic conditions. Cross-environmental testing will be a focal point, designed to assess our model's robustness in different scenarios. The goal is to verify the model's performance in an array of settings, thereby ensuring its practical applicability in real-world situations. Additionally, in our subsequent research endeavors, we plan to delve into the integration of our models within real-time systems, assessing the potential benefits and challenges.

The exploration of additional features will also be prioritized, with the intention of enhancing the model's predictive capabilities. Moreover, our commitment to improving the explain ability of AI systems remains steadfast, particularly as it pertains to their application in dynamic and unpredictable environments. The results of this study can provide guidance to transportation authorities, vehicle manufacturers, and developers on how to orchestrate the operation of connected vehicles in a more systematic manner. Additionally, they can assist drivers in better planning their routes and minimizing their wait times.

Author Contributions: Conceptualization, Y.K.J. and M.N.; methodology, Y.K.J. and M.N.; software, Y.K.J. and M.N.; validation, Y.K.J. and M.N.; formal analysis, Y.K.J. and M.N.; investigation, Y.K.J. and M.N.; resources, Y.K.J. and M.N.; data curation, Y.K.J. and M.N.; writing—original draft preparation, Y.K.J. and M.N.; writing—review and editing, Y.K.J. and M.N.; visualization, Y.K.J. and M.N.; supervision, Y.K.J. and M.N.; project administration, Y.K.J. and M.N.; funding acquisition, Y.K.J. and M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon reasonable request from the corresponding author. The data were accessed on 10 June 2024.

Acknowledgments: We acknowledge the dataset made available on Kaggle, which significantly enhanced the quality and depth of our research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Marcillo, P.; Tamayo-Urgilés, D.; Valdivieso Caraguay, Á.L.; Hernández-Álvarez, M. Security in V2I Communications: A Systematic Literature Review. *Sensors* **2022**, *22*, 9123. [CrossRef] [PubMed]
2. Han, B.; Peng, S.; Wu, C.; Wang, X.; Wang, B. LoRa-Based Physical Layer Key Generation for Secure V2V/V2I Communications. *Sensors* **2020**, *20*, 682. [CrossRef] [PubMed]
3. Jawad, Y.K.; Nitulescu, M. Transportation Systems for Intelligent Cities. In Proceedings of the 2023 24th International Carpathian Control Conference (ICCC), Miskolc-Szilvásvárad, Hungary, 12–14 June 2023; IEEE: New York, NY, USA, 2023. [CrossRef]
4. De Ville, B. Decision trees. *Wiley Interdiscip. Rev. Comput. Stat.* **2013**, *5*, 448–455. [CrossRef]
5. Zhang, Z. Introduction to machine learning: K-nearest neighbors. *Ann. Transl. Med.* **2016**, *4*, 218. [CrossRef] [PubMed]
6. Ye, H.; Liang, L.; Li, G.Y.; Kim, J.; Lu, L.; Wu, M. Machine learning for vehicular networks. *arXiv* **2017**, arXiv:1712.07143.
7. Petch, J.; Di, S.; Nelson, W. Opening the black box: The promise and limitations of explainable machine learning in cardiology. *Can. J. Cardiol.* **2022**, *38*, 204–213. [CrossRef]
8. Von Eschenbach, W.J. Transparency and the black box problem: Why we do not trust AI. *Philos. Technol.* **2021**, *34*, 1607–1622. [CrossRef]
9. Wischmeyer, T. Artificial Intelligence and Transparency: Opening the Black Box. In *Regulating Artificial Intelligence*; Springer: Cham, Switzerland, 2019. [CrossRef]
10. Rai, A. Explainable AI: From black box to glass box. *J. Acad. Mark. Sci.* **2020**, *48*, 137–141. [CrossRef]
11. Krawczyk, B.; Minku, L.L.; Gama, J.; Stefanowski, J.; Woźniak, M. Ensemble learning for data stream analysis: A survey. *Inf. Fusion* **2017**, *37*, 132–156. [CrossRef]

12. Mienye, I.D.; Sun, Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access* **2022**, *10*, 99129–99149. [CrossRef]
13. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2020**, *14*, 241–258. [CrossRef]
14. Jung, H.G.; Lim, K.T.; Shin, D.K.; Yoon, S.H.; Jin, S.K.; Jang, S.H.; Kwak, J.M. Reliability verification procedure of secured V2X communication for autonomous cooperation driving. In Proceedings of the 2018 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Republic of Korea, 17–19 October 2018; pp. 1356–1360. [CrossRef]
15. Jones, L. Driverless when and cars: Where? [Automotive Autonomes Vehicles]. *Eng. Technol.* **2017**, *12*, 36–40. [CrossRef]
16. Hussain, R.; Zeadally, S. Autonomous cars: Research results, issues and future challenges. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 1275–1313. [CrossRef]
17. Amiri, P.A.D.; Pierre, S. An Ensemble-Based Machine Learning Model for Forecasting Network Traffic in VANET. *IEEE Access* **2023**, *11*, 22855–22870. [CrossRef]
18. Lattanzi, E.; Freschi, V. Machine learning techniques to identify unsafe driving behavior by means of in-vehicle sensor data. *Expert Syst. Appl.* **2021**, *176*, 114818. [CrossRef]
19. Alvarez Coello, D.; Klotz, B.; Wilms, D.; Fejji, S.; Gómez, J.M.; Troncy, R. Modeling dangerous driving events based on in-vehicle data using random forest and recurrent neural network. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019.
20. Zhan, H.; Gomes, G.; Li, X.S.; Madduri, K.; Sim, A.; Wu, K. Consensus Ensemble System for Traffic Flow Prediction. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 3903–3914. [CrossRef]
21. Wang, W.; Xi, J. A rapid pattern-recognition method for driving styles using clustering-based support vector machines. In Proceedings of the American Control Conference (ACC), Boston, MA, USA, 6–8 July 2016; pp. 5270–5275.
22. Osman, O.A.; Hajij, M.; Bakhit, P.R.; Ishak, S. Prediction of near-crashes from observed vehicle kinematics using machine learning. *Transp. Res. Rec.* **2019**, *2673*, 463–473. [CrossRef]
23. Moreira, M.L.; Farah, H. On developing a driver identification methodology using in-vehicle data recorders. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 2387–2396. [CrossRef]
24. Ghadge, M.; Pandey, D.; Kalbande, D. Machine learning approach for predicting bumps on road. In Proceedings of the International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Davangere, India, 29–31 October 2015; pp. 481–485.
25. Dhiman, A.; Klette, R. Pothole detection using computer vision and learning. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 3536–3550. [CrossRef]
26. Kim, T.; Ryu, S.-K. Review and analysis of pothole detection methods. *J. Emerg. Trends Comput. Inf. Sci.* **2014**, *5*, 603–608.
27. Bernas, M.; Płaczek, B.; Korski, W.; Loska, P.; Smyła, J.; Szymała, P. A survey and comparison of low-cost sensing technologies for road traffic monitoring. *Sensors* **2018**, *18*, 3243. [CrossRef]
28. Martinelli, F.; Mercaldo, F.; Nardone, V.; Orlando, A.; Santone, A. Who's driving my car? A machine learning based approach to driver identification. In Proceedings of the 4th International Conference, ICISSP, Madeira, Portugal, 22–24 January 2018; pp. 367–372.
29. Martinelli, F.; Mercaldo, F.; Santone, A. Machine learning for driver detection through can bus. In Proceedings of the IEEE 91st Vehicular Technology Conference, Antwerp, Belgium, 25–28 May 2020; pp. 1–5.
30. Goh, C.C.; Kamarudin, L.M.; Zakaria, A.; Nishizaki, H.; Ramli, N.; Mao, X.; Syed Zakaria, S.M.M.; Kanagaraj, E.; Abdull Sukor, A.S.; Elham, M.F. Real-time in-vehicle air quality monitoring system using machine learning prediction algorithm. *Sensors* **2021**, *21*, 4956. [CrossRef] [PubMed]
31. Bai, R.; Chen, X.; Chen, Z.L.; Cui, T.; Gong, S.; He, W.; Jiang, X.; Jin, H.; Jin, J.; Kendall, G. Analytics and machine learning in vehicle routing research. *Int. J. Prod. Res.* **2023**, *61*, 4–30. [CrossRef]
32. Jawad, Y.K.; Nitulescu, M. Smart City Concepts and Urban Service Robots. In Proceedings of the SYROM 2022 & ROBOTICS, Iasi, Romania, 17–18 November 2022; Springer: Cham, Switzerland, 2023. [CrossRef]
33. Zhao, L.; Xu, T.; Zhang, Z.; Hao, Y. Lane-Changing Recognition of Urban Expressway Exit Using Natural Driving Data. *Appl. Sci.* **2022**, *12*, 9762. [CrossRef]
34. Al-refai, G.; Elmoaqet, H.; Ryalat, M. In-Vehicle Data for Predicting Road Conditions and Driving Style Using Machine Learning. *Appl. Sci.* **2022**, *12*, 8928. [CrossRef]
35. Volvo. Volvo City Safety Technology Guide. 2018. Available online: https://www.volvocarscincinnatieast.com/volvo-city-safety-technology-guide.htm (accessed on 28 August 2019).
36. Wu, C.H.; Ho, J.M.; Lee, D.T. Travel-time prediction with support vector regression. *IEEE Trans. Intell. Transp. Syst.* **2024**, *5*, 276–281. [CrossRef]
37. Reddy, K.K.; Kumar, B.A.; Vanajakshi, L. Bus travel time prediction under high variability conditions. *Curr. Sci.* **2016**, *111*, 700–711. [CrossRef]
38. Patnaik, J.; Chien, S.; Bladikas, A. Estimation of bus arrival times using APC data. *J. Public Transp.* **2004**, *7*, 1–20. [CrossRef]
39. Jeong, R.; Rilett, R. Bus arrival time prediction using artificial neural network model. In Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems, Washington, WA, USA, 3–6 October 2004.
40. Yu, K.; Peng, L.; Ding, X.; Zhang, F.; Chen, M. Prediction of instantaneous driving safety in emergency scenarios based on connected vehicle basic safety messages. *J. Intell. Connect. Veh.* **2019**, *2*, 78–90. [CrossRef]

41. Balid, W.; Tafish, H.; Refai, H.H. Intelligent vehicle counting and classification sensor for real-time traffic surveillance. *IEEE Trans. Intell. Transp. Syst.* **2017**, *19*, 1784–1794. [CrossRef]
42. Menegazzo, J.; von Wangenheim, A. Multi-Contextual and Multi-Aspect Analysis for Road Surface Type Classification through Inertial Sensors and Deep Learning. In Proceedings of the X Brazilian Symposium on Computing Systems Engineering (SBESC), Florianopolis, Brazil, 24–27 November 2020; pp. 1–8. [CrossRef]
43. Ibtissem, K.; Sami, F.; Souhayel, G. R-Secure: A system based on crowdsourcing platforms to improve road safety in the smart city. In Proceedings of the International Conference on Innovations in Intelligent Systems and Applications (INISTA), Biarritz, France, 8–12 August 2022.
44. Xu, Q.; Li, K.; Wang, J.; Yuan, Q.; Yang, Y.; Chu, W. The status, challenges, and trends: An interpretation of technology roadmap of intelligent and connected vehicles in China. *J. Intell. Connect. Veh.* **2020**, *5*, 1–7. [CrossRef]
45. Li, H.; Zhang, J.; Zhang, Z.; Huang, Z. Active Lane management for intelligent connected vehicles in weaving areas of urban expressway. *J. Intell. Connect. Veh.* **2021**, *4*, 52–67. [CrossRef]
46. Mao, S.; Xiao, G.; Lee, J.; Wang, L.; Wang, Z.; Huang, H. Safety effects of work zone advisory systems under the intelligent connected vehicle environment: A microsimulation approach. *J. Intell. Connect. Veh.* **2021**, *4*, 16–27. [CrossRef]
47. He, Z.; Chen, Y.; Zhang, H.; Zhang, D. WKN-OC: A new deep learning method for anomaly detection in intelligent vehicles. *IEEE Trans. Intell. Veh.* **2023**, *8*, 2162–2172. [CrossRef]
48. Paden, B.; Čáp, M.; Yong, S.Z.; Yershov, D.; Frazzoli, E. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Trans. Intell. Veh.* **2016**, *1*, 33–55. [CrossRef]
49. Jawad, Y.K.; Nitulescu, M. Controlling Traffic Congestion in a Residential Area via GLOSA Development. *Appl. Sci.* **2024**, *14*, 1474. [CrossRef]
50. Cai, Y.; Zhao, R.; Wang, H.; Chen, L.; Lian, Y.; Zhong, Y. CNN-LSTM Driving Style Classification Model Based on Driver Operation Time Series Data. *IEEE Access* **2023**, *11*, 16203–16212. [CrossRef]
51. Zhao, T.; He, J.; Lv, J.; Min, D.; Wei, Y. A Comprehensive Implementation of Road Surface Classification for Vehicle Driving Assistance: Dataset, Models, and Deployment. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 8361–8370. [CrossRef]
52. Ahmed, U.; Tu, R.; Xu, J.; Amirjamshidi, G.; Hatzopoulou, M.; Roorda, M.J. GPS-Based Traffic Conditions Classification Using Machine Learning Approaches. *Transp. Res. Rec.* **2022**, *2677*, 1445–1454. [CrossRef]