**Standards-related Regional Innovation and International Cooperation**
https://journals.zycentre.com/sriic

**ARTICLE**

# Machine Learning Digital Twin Applied to Hybrid Vehicle Emission Test—A Multimetric Evaluation Approach

***Natalia Miedviedieva*** [1] , ***Eduardo Tomanik*** [2*] , ***Ellen Rodrigues*** [3], ***Fernando Fusco Rovai*** [4,5]

[1] *Air Transportation Management Department, National University"Kyiv Aviation Institute", 03058 Kyiv, Ukraine*
[2] *Escola Politecnica, Universidade de Sao Paulo, Sao Paulo 05508-010, Brazil*
[3] *Independent Researcher, Farmington Hills, MI 48331, USA*
[4] *Volkswagen do Brasil, São Bernardo do Campo 09823-901, Brazil*
[5] *Department of Mechanical Engineering, Centro Universitário FEI, São Bernardo do Campo 09850-901, Brazil*

## ABSTRACT

This article proposes and illustrates a multimetric evaluation template for digital twins based on machine learning in critical engineering applications using an example as a specific testbed for discussing unified assessment principles. We analyze this concept through the lens of a particular and relevant example: the performance of a hybrid vehicle during homologation tests for transient emissions. In this scenario, ML models must not only optimize efficiency but also ensure strict compliance with environmental regulations in dynamic, operating modes. The case illustrates the complexity that arises when attempting to unify requirements for accuracy, fault tolerance, adaptability, and regulatory compliance, providing a framework for exploring how a unified evaluation system can lead to a more consistent and reliable integration of ML into critical systems. Test data of emission tests on a Hybrid vehicle was used to train a Random Forest model. Different sets of input parameters illustrate some of the capabilities and limitations of using AI. Shapley values were used to discuss some of the AI model characteristics and limitations. Using as input parameters only Vehicle speed, Acceleration, and Battery State of Charge (SoC) allowed the digital twin to achieve R2 0.80. Inclusion of Internal Combustion Engine oil temperature increased the model R2 to 0.97 and curiously changed the ranking of the other input parameters. SoC, which

*CORRESPONDING AUTHOR:
Eduardo Tomanik, Escola Politecnica, Universidade de Sao Paulo, Sao Paulo 05508-010, Brazil; Email: Eduardo.tomanik@gmail.com

ARTICLE INFO
Received: 14 June 2025 | Revised: 6 August 2025 | Accepted: 13 August 2025| Published Online: 20 August 2025
DOI: https://doi.org/10.63385/sriic.v1i2.348

CITATION
Miedviedieva, N., Tomanik, E., Rodrigues, E., et al., 2025. Machine Learning Digital Twin Applied to Hybrid Vehicle Emission Test—A Multimetric Evaluation Approach. Standards-related Regional Innovation and International Cooperation. 1(2): 55–78.
DOI: https://doi.org/10.63385/sriic.v1i2.348

was the most influential in previous cases, went down to almost negligible impact on the model results. Despite reaching a quite high R2 of 0.97, the model can miss important aspects of the physical system without Human expert supervision.
***Keywords:*** Artificial Intelligence; Machine Learning; Dynamic Vehicle Simulation; Vehicle Electrification; Multimetric Evaluation

# 1. Introduction

Artificial Intelligence, and in particular machine learning (ML) tools, are becoming a cornerstone of innovation; their integration into critical areas, from healthcare to autonomous systems, promises significant breakthroughs. However, this integration also creates serious problems, as the consequences of failures in such systems can be catastrophic. Despite the growing reliance on ML, current methods of evaluating effectiveness often remain fragmented and specific to particular domains. Different application domains prioritize various performance metrics, validation procedures, and reporting standards, complicating the comparison of models, certification, and the establishment of trust across domains. Although several normative and methodological frameworks have been proposed, such as ISO/IEC standards[1], the NIST AI risk management framework[2], or reporting recommendations for specific domains, there is still no widely accepted operational template that integrates general principles of assessment with specific, applied practices for particular fields.

The challenges faced by developers and regulators of ML are significantly complicated by industry-specific requirements. Each critical sector—whether it be autonomous driving, medical diagnostics, financial forecasting, or energy management—has its own unique set of success criteria, error margins, and regulatory obligations. This specialization, although necessary for adapting machine learning to specific needs, is characterized by significant discrepancies in evaluation methods. For example, in the medical field, the effectiveness of a model is determined by sensitivity, specificity of the diagnosis, and the degree of explainability for clinicians, while in the energy sector, the main criteria are system stability and accuracy of demand forecasting. Such diversity of criteria complicates the comparability of models across domains, making it impossible to develop universal metrics for reliability, safety, and fault tolerance.

In the automotive industry, and particularly in emissions testing of hybrid vehicles under transient conditions, the issues mentioned before are especially pronounced. Hybrid powertrains operate in high-dynamic modes, balancing the operation of internal combustion engines, electric motors, energy regeneration, and battery management. Machine learning-based digital twins have emerged as a promising tool for supporting the analysis and prediction of emissions in such conditions, offering scalability and flexibility compared to traditional calibration and testing methods. However, without a structured and transparent evaluation system, high predictive accuracy may conceal limited physical validity, poor generalizability, or hidden data dependencies.

Consequently, there is a need to create a unified, flexible evaluation template that can be adapted to specific contexts. Multimetric templates tailored to specific subject areas can streamline the development and reporting of machine learning systems while maintaining a consistent level of trust, ethics, and regulatory compliance in critical applications. The lack of unification, in turn, creates the risk of forming disparate, unverified systems that could undermine trust in machine learning and reduce its transformational potential.

# 2. Literature Review

## 2.1. Overview and the Need for Standardized Evaluation Templates in Machine Learning

Machine learning models deployed in critical domains should be evaluated along multiple dimensions, including predictive performance, robustness to data shifts, data quality, explainability, and, where appropriate, fairness and ethical impact. Such a multidimensional aspect demands flexible evaluation templates that combine general principles with domain-specific instantiations. A range of quantitative indicators covering various aspects of model performance, such

as accuracy, precision, recall, F1 score, area under the ROC curve (AUC), calibration, and generalization during data splitting (e.g., cross-validation, holdout)[3]. For critical areas, additional metrics tailored to the specific needs of the industry should be included (e.g., risk metrics for health or fault diagnosis). To ensure a comprehensive comparison of classifiers, multimetric frameworks have been proposed that simultaneously evaluate models across several indicators[4]. The F1 score is the harmonic mean of these two metrics, making it particularly useful in situations where a compromise is needed between a model's ability to avoid false positives and false negatives. The F1 score is a robust criterion when working with imbalanced datasets, where one class significantly outweighs the other. Unlike accuracy, the F1 score is not distorted by class imbalance and better reflects the true effectiveness of classification. The ROC curve illustrates the relationship between the True Positive Rate (TPR, or Recall) and the False Positive Rate (FPR). The ROC curve allows for a visual assessment of the stability and sensitivity of the model at various threshold levels, as well as the simultaneous comparison of multiple models. The ROC is independent of a specific classification threshold, in contrast to the F1-score, which is based on a fixed threshold for decision-making. AUC reflects the probability that the model correctly distinguishes between a random positive and a negative example. AUC is independent of the chosen threshold and provides a global assessment of the model's discriminative ability. While ROC offers a visual representation, AUC provides a quantitative evaluation—the closer to 1, the more effective the model; a value of 0.5 corresponds to a random classifier. LIME creates a simplified, interpretable model (typically linear) around a specific instance by artificially varying its input data. LIME provides a comprehensible local explanation—demonstrating which particular features have the most significant impact on an individual prediction. It is focused on the local interpretation of individual cases, whereas SHAP allows for global generalizations across the entire model. SHAP is based on Shapley values from game theory, distributing the "contribution" of each feature to the model's prediction in a fair and mathematically justified manner. SHAP provides both local and global interpretation—meaning it explains individual predictions as well as the overall structure of feature importance. Unlike LIME, SHAP guarantees consistency and fairness in feature importance assessments due to its rigorous theoretical foundation (Shapley values).

Resilience to distributional shifts (covariate shifts) and competitive examples is another important measure. A new fundamental framework for assessing reliability emerges, which goes beyond classical accuracy metrics and employs theoretically grounded metrics such as Posterior Agreement, offering a substantive analysis of distributional changes prevalent in critical applications[5]. Methods of explainable artificial intelligence (XAI) are increasingly recognized as a key element in the evaluation of machine learning models, particularly in critical environments such as healthcare, finance, and cybersecurity. Various post-hoc interpretation methods (LIME, SHAP, counterfactual explanations) assist in understanding the decisions of black-box models and are essential for trust and regulatory compliance[6–8]. Probabilistic assessment and decision lists offer alternatives based on uncertainty that are more interpretable and suitable for decisions critically important to safety[9].

The data utilized for training and testing machine learning models must undergo rigorous engineering processes to ensure quality, completeness, consistency, and representativeness, as these factors influence the validity and reliability of the model[10]. Avoiding pitfalls such as data leakage, which leads to overly optimistic estimates, is critically important and should be standard practice[11]. Similarly, standards must consider bias and fairness considerations to prevent the perpetuation of social inequality and discriminatory outcomes[12].

Considering the diversity of domains, a unified structure should incorporate methods that allow for the adaptation and generalization of domains, including transfer learning and federated or decentralized learning to work with heterogeneous datasets and privacy constraints[11, 13]. Basic models are promising due to their broad coverage and adaptability, but they require evaluative metrics aligned with domain performance, security, privacy, and relevance[14].

Standardization in the development and reporting of models is crucial for promoting reproducibility and comparability. Clinical and industry standards, such as TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) and PROBAST (Prediction Model Risk of Bias Assessment Tool), can be adapted to machine learning settings for proper reporting and bias

risk assessment[14, 15]. These checklists guide evaluation, enabling stakeholders who are not experts to critically assess machine learning models and ensure their accuracy and reliability.

Data augmentation methods can address the issue of data scarcity, which is a common limitation in critical applications. Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and diversity-based perspectives enhance data efficacy and model reliability[16]. However, it is critically important to assess and mitigate potential biases introduced by data augmentation, especially when augmented datasets are utilized for both training and testing.

Review[17] shows that even in the financial market domain, researchers are moving toward evaluation patterns built around an end-to-end workflow + set of metrics + reproducibility recommendations, which is consistent with our idea of a multi-metric evaluation template for digital twins that combines general ML evaluation principles with domain-specific interpretation for hybrid vehicles.

Ethical and socio-technical considerations are an integral part of the evaluation of machine learning models in high-stakes domains. Insights from governance and incentive structures in decentralized digital ecosystems, such as blockchain-based systems[13], together with frameworks for fairness and bias mitigation in AI[12], can inform the design of collaborative evaluation practices that explicitly account for ethics, privacy, and societal impact.

Recent research on the evaluation of machine learning (ML) in safety-critical domains emphasizes that no single performance metric is sufficient to characterize model behavior, particularly when regulatory compliance and human oversight are required. Instead, multimetric templates are recommended, combining accuracy, calibration, robustness to distributional shifts, and uncertainty assessment, to encompass various failure modes of the model and to support risk-aware decision-making. Concurrently, several works on AI explainability and data quality assessment argue that evaluation should jointly consider the quality of the underlying dataset, the stability of explanations, and potential fairness issues, rather than treating them as separate checks applied in isolation.

In the automotive and transportation sectors, research on digital twins and predictive maintenance increasingly relies on ensemble models, such as random forests or gradient-boosted trees, along with local explanation methods (e.g., SHAP) to interpret the behavior and emissions of the powertrain. These contributions typically optimize models for specific tasks or driving cycles and report a limited set of metrics (e.g., $R^2$ and RMSE), while aspects such as data quality indicators or systematic robustness checks are discussed only qualitatively. This article builds upon this body of work, offering a structured evaluation template with multiple metrics and illustrating it through a specific case study of hybrid vehicle emissions testing, clearly linking general metrics and explanation tools with domain-specific choices of input signals and operating modes.

As discussed before, there is a need for a comprehensive toolkit for their evaluation to overcome the fragmentation of approaches and ensure progress in the development of reliable and ethical machine learning systems in critical domains. In this context, **Tables 1–4** present a systematic review of the key performance metrics, data quality indicators, methods for ensuring explanations and fairness, as well as correlation metrics. The tables outline advantages and limitations, and most importantly, potential areas for further enhancement and contextual application, which could serve as a foundation for the development of unified, adaptive, and integrated frameworks for evaluation.

**Table 1.** Data Quality Indicators.

| Metric | Advantages | Disadvantages | Potential Improvements |
|---|---|---|---|
| Missing values % | Simple Data Quality Control | Does not affect the model. | Employ imputation strategies, data cleaning, or feature engineering to address missing values; collect more data whenever feasible. |
| Label noise | It is important in classification. | It is challenging to accurately assess. | Use robust learning algorithms, anomaly detection methods for labels, and ensembles to mitigate the impact of noise; consider soft labels. |
| Accumulated Error % | Helps to assess accuracy | It may be subjective. | Use of outlier detection methods (e.g., Isolation Forest, Local Outlier Factor), robust algorithms, data transformation, and analysis of error causes. |

**Table 2.** Data Clustering Methods.

| Metric | Advantages | Disadvantages | Potential Improvements |
|---|---|---|---|
| Silhouette Score | Measures the density of clusters and their separation; values range from −1 to 1. | It can be misleading due to the aggregation of incorrect shapes or varying densities. | Combine with visual analysis (e.g., *t*-SNE, UMAP visualization). Consider other internal validation metrics (e.g., Calinski-Harabasz, Davies-Bouldin) or external ones, if they are true. |
| Davies-Bouldin Index | A lower value means better clusters (more compact and better separated). | Does not always correlate with the intuitive understanding of "correct" clustering; may be sensitive to the shape of clusters. | Combine with visual analysis. Experiment with various clustering algorithms, as this metric may favor certain cluster structures. If possible, consider external validation metrics. |

**Table 3.** Classification Metrics.

| Metric | Advantages | Disadvantages | Potential Improvements |
|---|---|---|---|
| Accuracy | Ease of interpretation, general understanding of quality | Sensitivity to class imbalance | Use in conjunction with Precision, Recall, F1-Score, or ROC AUC, especially for imbalanced datasets. Consider the cost of errors. |
| Precision | It is crucial to consider the implications of a false positive result (medications, safety) at a high cost. | Ignores false negative results | Evaluate alongside the Recall or F1-Score for a comprehensive understanding. Assess the impact of false negatives and false positives on the business. |
| Recall | It is important for the price of high throughput (false negative result) | Ignores false positive results | Evaluate in conjunction with Precision or F1-Score. Determine the impact of false negatives and false positives on the business. |
| F1-score | Balance in tasks with imbalanced classes | It is challenging to interpret out of context. | Consider the F-beta score (Fβ score) when one type of error (precision or recall) is more critical. Supplement the ROC AUC for overall discriminative ability. |
| AUC-ROC | Good evaluation of the quality at various thresholds. | Complex interpretation in multiclass tasks | Supplement with precision and recall curves (PRC AUC) for highly imbalanced datasets. Establish decision thresholds based on specific business costs. |

**Table 4.** Metrics for Evaluating Machine Learning Models.

| Metric Type | Metric | Advantages | Disadvantages | Potential Improvements |
|---|---|---|---|---|
| Regression | MSE | Order for large deviations | Highly sensitive to localized errors | Use of RMSE for enhanced interpretation. Consider the significant probability of substantial anomalous variation (MAE) for critical issues. Investigate the application of robust regression methods. |
| | RMSE | Easier to interpret | Still sensitive to localized errors | Use MAE if the tolerance to deviations is critical. Analyze the residuals graphs to understand the distribution of errors. |
| | TIIS/ITIS | Radiation-resistant | Ignores large deviations | Use in conjunction with RMSE to understand the impact of larger errors. Consider Huber loss as a smooth and robust alternative during training. |
| | $R^2$ | A good interpretation of model quality | Can be negative, depending on the range of data | Use the adjusted coefficient of determination ($R^2$−) to penalize excessive predictors. Examine residual plots for linearity, homoscedasticity, and independence of errors. |
| Correlational | Pearson | Simplicity, a well-known metric | Does not account for non-linearity | Visualize data (e.g., scatter plots) before interpretation. Consider other coefficients if the relationship is non-linear or if there are outliers. Check for normality. |

**Table 4.** *Cont.*

| Metric Type | Metric | Advantages | Disadvantages | Potential Improvements |
|---|---|---|---|---|
| Correlational | Spearman | Resistance to monotonic nonlinear relationships | Less interpretable in absolute values | Use when the relationship is monotonic but not necessarily linear, or when the data contain outliers. Well-suited for ordinal data. |
| | Tau Kendall | More reliable for small samples | Computationally more expensive | Use for ordinal data or when a reliable measure of monotonicity is required. Often employed as an alternative to Spearman's method, especially with a large number of tied ranks. |
| Stability and Generalizability | Stability of Cross-Validation | Provides insight into generalizability | Computationally complex for large datasets | Use strategic sampling (stratified coefficient of variation), increase number of folds, repeated coefficient of variation to reduce the variance of the estimate and obtain more reliable confidence intervals. |
| | External effectiveness testing | Realistic assessment | Data often absent | Regularly update external test data; monitor model performance post-deployment (MLOps); adapt domain when data distribution changes. |
| Interpretation and Reliability | Consistency of SHAP/LIME | Ensures Transparency | Resource-Intensive | Develop metrics for quantitative assessment of explanation stability; visual analysis of explanations for clusters of similar instances; use ensemble explanatory methods; check for contradictions in explanations. |
| | Indicators of fairness (e.g., demographic parity) | Important for critically essential areas | More group data is needed | Establish a clear definition of fairness for each domain; employ bias reduction methods (preprocessing, in-process treatment, post-processing); regular audit of model fairness; engage subject matter experts and representatives from the groups affected by the model. |

In the field of machine learning and artificial intelligence, standards play a crucial role in ensuring reliability, safety, transparency, and ethics. Although there is no universal "single standard" for machine learning due to its rapid development and diversity of applications, there are a number of initiatives, recommendations, and regulatory frameworks that form the foundation for responsible development and implementation. One of the first intergovernmental standards for responsible AI is the OECD principles on AI, adopted by OECD member countries. They encompass inclusive growth, sustainable development, human values, transparency, accountability, safety, and privacy. The AI Risk Management Framework (AI RMF), developed by the National Institute of Standards and Technology (NIST) in the United States, serves as a voluntary guide for managing AI risks, including bias, safety, and transparency. International standard ISO/IEC 23053:2022 "Framework for Artificial Intelligence (AI) Systems Utilizing Machine Learning (ML)"[1]: Describes the architecture and components of AI systems that employ ML, providing a foundation for their understanding and interaction. The European Union's Artificial Intelligence Act—This legislative proposal represents the world's first attempt to comprehensively regulate artificial intelligence by classifying systems according to their risk level and establishing stringent requirements for "high-risk" systems, which are often based on machine learning.

The analysis of existing machine learning standards underscores the necessity for a comprehensive evaluation of performance, explanations, data quality, domain adaptation, standardized reporting, and ethical considerations. Therefore, there is a need not for a single universal standard that encompasses all possible applications of machine learning (ML), but for evaluation templates that take into account the specifics of the subject area, yet are structured and can be developed under specific critical conditions. In this case, we focus on one such instance—the assessment of transient emissions from hybrid vehicles based on digital twins—and use it as a specific testing ground for discussing unified evaluation principles.

**Tables 1–4** provide a comprehensive overview of assessment tools that are relevant to numerous significant ma-

chine learning programs, including data quality metrics, clustering and classification metrics, regression and correlation metrics, as well as criteria for interpretability and fairness. This general overview is intended as a toolkit: various subject areas will develop only those elements that are meaningful for their specific tasks, data structures, and risk profiles.

The integration of machine learning (ML) systems in the performance assessment of hybrid vehicles, particularly during transient emissions testing, represents a critical advancement in the field of sustainable transport design. Transient emissions testing simulates real-world driving conditions, such as those stipulated by regulations concerning real driving emissions (RDE). Hybrid powertrains must balance fuel efficiency, reduction of pollutant emissions, and energy management under varying load and speed conditions. Traditional calibration methods for these systems often rely on hardware-in-the-loop (HiL) testing or portable emissions measurement systems (PEMS), which are resource-intensive and limited in scalability.

In this paper, we operationalize a targeted subset of the proposed multimetric by:

- From the measurement of data quality (**Table 1**), we apply handling of missing values and basic screening for outliers in variables related to emissions before model training.
- From regression and correlation metrics (**Tables 3** and **4**), we utilized $R^2$ and MSE as primary performance indicators, supplemented by Pearson and Spearman coefficients to characterize the relationships between input variables (vehicle speed, acceleration, Battery State of Charge, temperatures) and the target power share. To evaluate the model's capability to be used in different test cycles, we also used the accumulated error of the predicted output, $CO_2$ emissions, at the end of the test cycle, since such a value is used for vehicle homologation.
- For interpretability and reliability (**Table 4**), we created feature attribution based on SHAP to analyze how the model distributes importance among the input data and how the introduction of additional input parameters, such as engine oil temperature, affects the model results.

Clustering metrics were not utilized, as the case study was a supervised regression rather than unsupervised pattern detection, and classification metrics (accuracy, precision, recall, F1) are not applicable to the continuous power distribution target. Other elements in **Tables 1–4**, such as classification metrics, fairness indicators, or formal bias audits, are included in the overall template for critical machine learning systems but were not explicitly applied in this case study of hybrid vehicles.

The research exemplifies the transition from a general framework to a specific case of emission tests of a hybrid vehicle, discussing how the proposed multimetric template can be tailored to a particular engineering domain, without claiming universal applicability to all potential critical systems of machine learning.

## 2.2. Hybrid Vehicles and Emission Test

A Multimetric evaluation template for digital twins that incorporates standardization methods addresses the discussed issues by enhancing forecasting accuracy, reducing uncertainty, and ensuring the reliability of emission predictions. This is particularly relevant as the transportation sector accounts for approximately 25% of global greenhouse gas emissions, underscoring the necessity for robust machine learning models to optimize hybrid vehicles to minimize carbon footprints and comply with regulatory requirements[18]. By combining an internal combustion engine (ICE) with one (or more) electric motors (EM), hybrid vehicles possess smaller electric batteries, are less expensive, exhibit reduced dependence on infrastructure, present an alternative for decreasing $CO_2$ emissions, and occupy a significant market share globally, see **Figure 1**.

The development of a unified machine learning (ML) system for testing emissions of hybrid vehicles typically involves modeling approaches that combine data-driven methods with physical modeling. An important feature of ML algorithms is their ability to dynamically improve or "learn," which occurs concurrently with the increase in the volume of available data. Currently, the primary machine learning algorithms commonly employed in the testing of hybrid vehicles for transient emissions of pollutants are as follows, see **Figure 2**.

The selection of a specific algorithm depends on the volume and quality of available data, computational resources, accuracy requirements, interpretability, and model speed. Hybrid approaches are often employed, or multiple algorithms are compared to achieve optimal results.

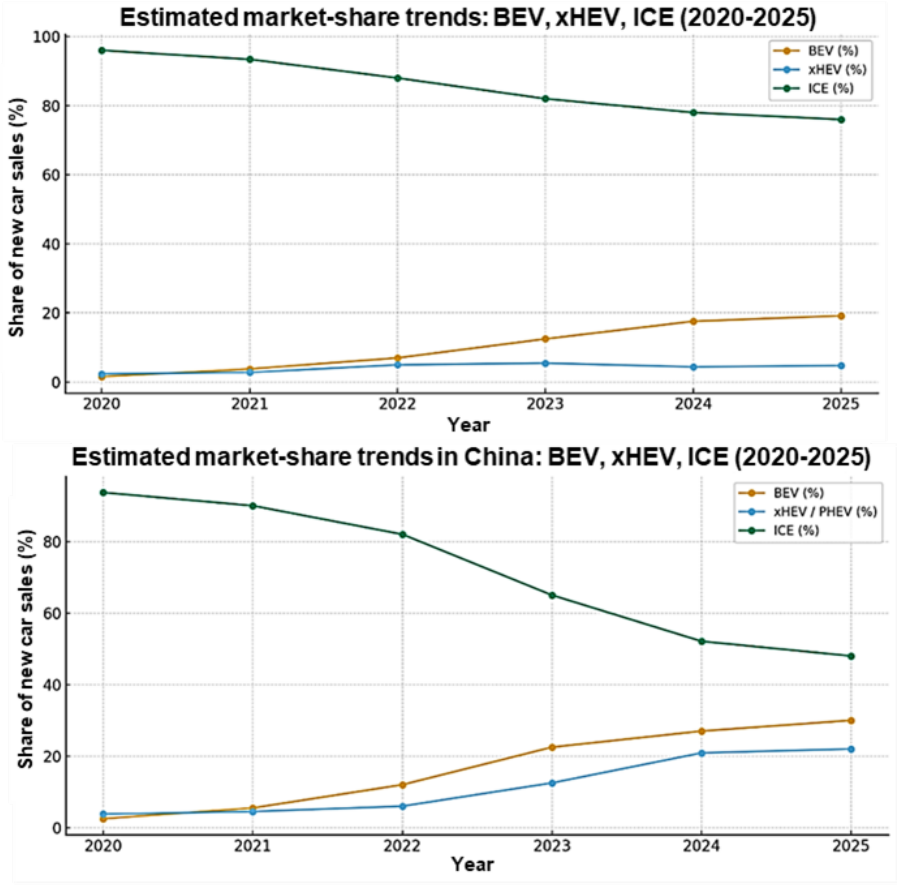


**Figure 1.** Market Share of Hybrid Vehicles.

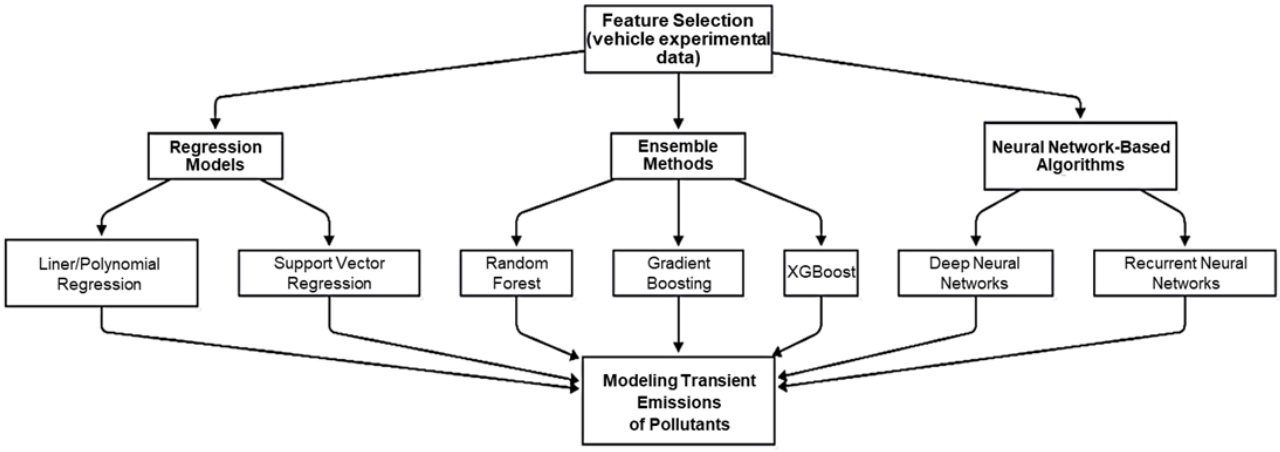


**Figure 2.** The Workflow for Machine Learning Systems.

Source: Adapted from Tomanik et al. and Ganesh and Xu [19, 20].

For effective modeling and forecasting of complex dynamic processes, a variety of algorithms are proposed that combine different approaches to data processing—from neural architectures to ensemble statistical models. Such algorithms include Recurrent Neural Networks (RNN), Long Short-Term Memory networks (LSTM), and Gated Recurrent Units (GRU), as well as ensemble methods such as Extreme Gradient Boosting (XGBoost), and statistical approaches represented by Support Vector Regression (SVR).

Algorithms such as RNN, LSTM, GRU, XGBoost, and

SVR represent various approaches to modeling and forecasting, each with its own advantages and specific data processing requirements. RNN (Recurrent Neural Network) effectively captures temporal dependencies in sequential data by retaining information about previous states.

However, the mentioned algorithms have limitations in reproducing long-term dependencies due to the "vanishing gradient" problem. LSTM (Long Short-Term Memory) addresses this limitation through memory cells and information flow control mechanisms. Stability and the ability to model long-term temporal dependencies are its strengths, making it suitable for forecasting complex dynamic processes. LSTM (Long Short-Term Memory) addresses this limitation through memory cells and information flow control mechanisms. Stability and the ability to model long-term temporal dependencies are its strengths, making it suitable for forecasting complex dynamic processes. Unlike neural networks, XGBoost (Extreme Gradient Boosting) implements an ensemble approach based on decision trees. Its key advantages include high accuracy, resistance to overfitting, and interpretability. The algorithm is particularly effective for tabular data and tasks of classification or regression. The SVR (Support Vector Regression) algorithm, based on the theory of support vectors, effectively models nonlinear dependencies using kernel functions. It is characterized by high accuracy on small samples and robustness to outliers, although it is less scalable for large datasets.

After the creation of a machine learning model, the process typically unfolds in several sequential stages, see **Figure 3**:

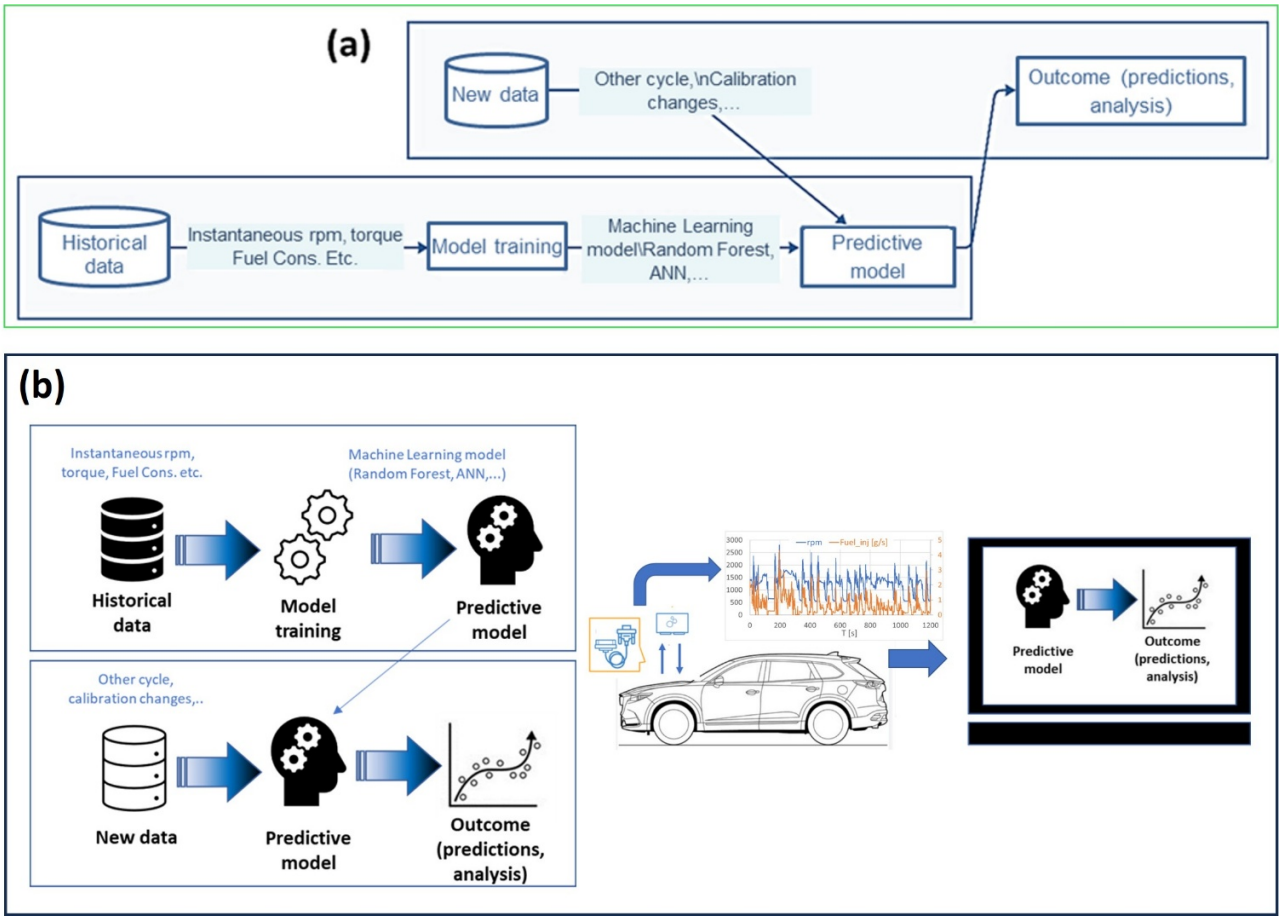


**Figure 3.** Stages of Creating a Machine Learning Model: (**a**) General; (**b**) As Used in This Work.

In this work, a standardized dynamometer highway driving cycle (HWFET) was used as example to assess the Power Share (PS), i.e., the instantaneous power ratio of the electrical motor to the total power demanded by the vehicle, which points out how much power from the internal combustion engine (ICE) is necessary to complement the electric

power of a Hybrid Vehicle.

The hybrid Vehicle PS is so significant for emissions results that it is covered by the test standard SAE J1711 [21], that determines the vehicle in-cycle measurement in both conditions: Charge Deplete (CD), in which the vehicle operates in electric mode (except if ICE is demanded to follow the cycle speed profile), and Charge Sustain (CS), in which the battery State of Charge (SoC) should be kept inside a determined range. Vehicle emissions and fuel economy are considerably different in these two operating modes, and this difference is strongly dependent on PS. The higher the electric motor power, the higher the energy recovery during vehicle deceleration and consequently the lower the dependence on ICE, which has considerably lower efficiency than an electric motor. Additionally, the emissions ($CO_2$ and pollutants) shared between CD and CS modes depend on the vehicle's electric range, which is directly determined by battery capacity. Higher electric ranges result in higher charge depletion influence on final results. The PS influence is already observed in Real Drive Emissions (RDE) tests, which cover urban, rural, and road conditions [22, 23].

Utilizing data from the 2021 Toyota RAV4 Prime provided by the Argonne National Laboratory (ANL), a random forest model was trained and validated to predict the instantaneous Power Share based on instantaneous driving parameters such as vehicle speed, acceleration, and battery state of charge.

# 3. Materials and Methods

Based on the multimetric evaluation template described in Section 2.2, this thematic study focuses on a random forest model that predicts the instantaneous power distribution between the electric motor and the internal combustion engine in a plug-in hybrid vehicle during a standardized emissions test on a dynamometer. In this section, we outline the dataset, the selection of input variables, as well as specific data quality checks, performance metrics, and explanatory tools applied in this case. A machine learning, Random Forest, model was selected for its good balance of accuracy, simplicity, and computer cost. Random Forest is a common machine learning meta-estimator that fits a number of decision tree regressors on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. See more details in Scikit-learn developers [24].

Random Forest has the advantage of working well without the need to normalize the data. In the code used in this work, the best model is defined by the one with the highest $R^2$. The AI model is an upgrade of the one used in previous author works [25, 26], and it was written in Python, using libraries Pandas, Numpy, Seaborn, and Scikit-Learn, and runs in the Jupyter Lab inside the Anaconda Software, see **Figure 4**.

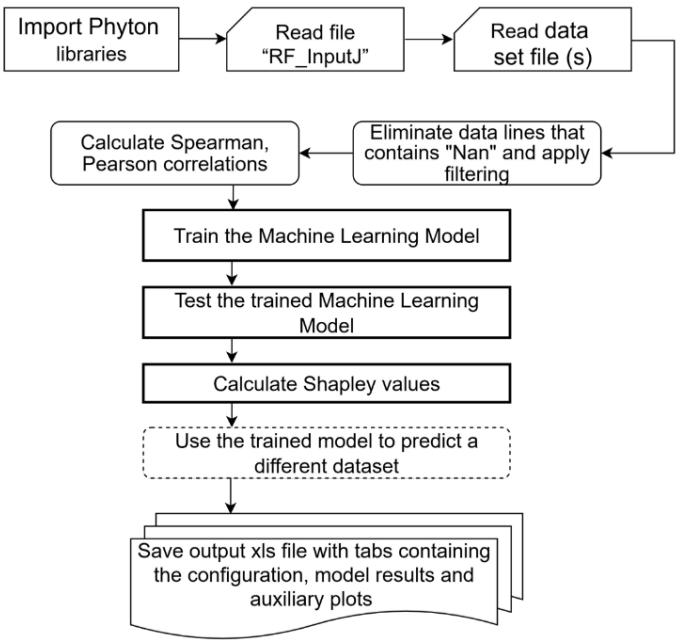


**Figure 4.** Code Flowchart.

To reduce the inherent risks associated with hard-coding configuration parameters directly into the script and to facilitate a more robust organizational structure, the code reads a dedicated input file called "RF_InputJ" where the user defines the dataset file name, input parameters, etc. This input file serves as a comprehensive configuration manifest containing detailed information such as the specific name of the data set being processed, the ratio of training/test, the column numbers that contain the parameters, and other relevant processing directives, see **Figure 5**. This approach centralizes all critical operational settings, thereby eliminating the need to directly modify the code when configuring parameters for different runs or data sets. This greatly reduces the potential for human error and increases the maintainability and reusability of the code base.

After successful execution, for better documentation, the code stores a copy of the input file and all generated results. This includes the primary analytical results, key visualizations and graphs summarizing the findings, and, most importantly, a complete record of the specific conditions under which the run was performed. This careful recording of run conditions is vital to ensure full reproducibility, facilitate accurate tracking of results, and enable effective comparison between different experimental runs or model iterations. Such organized management of results not only simplifies subsequent analysis but also supports strict version control and auditing of research results.

The code also calculates the SHAP (SHapley Additive exPlanations) values, which assign to each variable an importance (SHAP value), indicating how much it contributes to increasing or decreasing the model's prediction for a specific observation[27].

The code was applied to datasets of vehicle tests carried out and published by the Argonne National Lab[28] in a Hybrid vehicle, the 2021 Toyota RAV4 PRIME. **Figure 6** illustrates the key characteristics of the vehicle. **Figure 7** and **Table 5**, reproduced from the Toyota manual, overview the operation of the EV and Hybrid driving modes.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| File Name | RV4 Hwy 4013.xlsx | | # dataset used for training | | | | | | | | |
| Input Columns | 2,5,6,117,258 | | # indexes of the columns containing the input parameters | | | | | | | | |
| Output Column | 106 | | # column index containing the target parameter | | | | | | | | |
| Train/Test Split | 0.7 | | # ratio of training | | | | | | | | |
| Random State | 42 | | # seed for the Random initialization | | **column indexes:** | | | | | | |
| RF_n_estimators | 10-100:2 | | # range for the optimization loop (start-end: step) | | 2 | km/h | | | | | |
| RF_max_depth | None | | # Number or None | | 4 | torq | | | | | |
| RF_random_state | 42 | | # Model seed | | 5 | Acc | | | | | |
| Filter Condition | | | # optional: filters applied to the dataset | | 6 | PS | Power Split (eM/Total) | | | | |
| SHAP Dependence Feature | Power Split (eM/Total) | | # feature used in Shap dependence plot | | 77 | T_oil | Engine_Oil_Dipstick_Temp[C] | | | | |
| | | | | | 101 | THC | AMA_Dilute_THC[mg/s] | | | | |
| | | | | | 102 | CH4 | AMA_Dilute_CH4[mg/s] | | | | |
| | | | | | 103 | NOx | AMA_Dilute_NOx[mg/s] | | | | |
| | | | | | 106 | CO2 | AMA_Dilute_CO2[mg/s] | | | | |
| | | | | | 117 | | pedal_accel_pos_CAN2_per | | | | |
| | | | | | 253 | | HVBatt_refrigerant duct1_outlet_temp_BECM[C] | | | | |
| | | | | | 258 | SoC | | | | | |
| | | | | | 303 | Catal y T | | | | | |

**Figure 5.** Input File Defining the Parameters of Each Case Run.

Note: The indices in the "input columns" identify the parameters ("km/h", "Acc", etc.). The other lines are the "target" to be predicted ($CO_2$ in the example) and the applied code options.

Engine: 2.5-liter, I4, DI and PFI
177 hp @ 6,000 rpm; 165 lb.-ft. @3,600 rpm
Electric Powertrain: Plug-in Hybrid (42 mi EV Range)
AWD w/Permanent Magnet Rear Motor
MG1/MG2: 134 kW, 270 Nm
MGR: 40 kW, 121 Nm
Battery: Li-Ion 355. 2V, 51 Ah (18.1 kWh)
Transmission: Electronic Continuously Variable
Final Reduction Ratio (Front / Rear): 3.412 / 10.718

**Figure 6.** Key Characteristics of Toyota RAV4 2021 Tested at ANL[29].

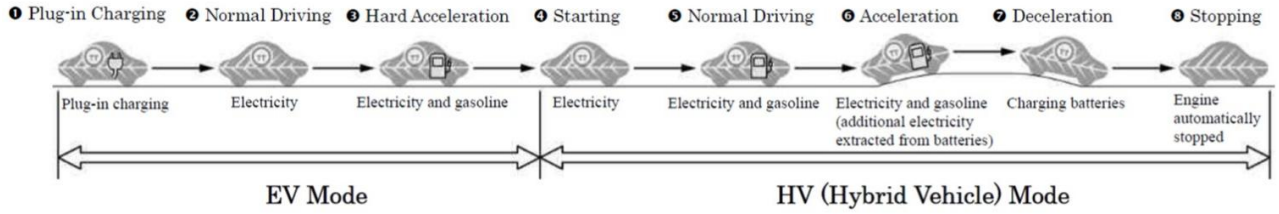


**Figure 7.** Toyota Driving Modes.

Note: Reproduced from the Toyota manual [29].

**Table 5.** Toyota RAV 4 Vehicle Driving Modes.

| 1. EV (Electric Vehicle) Mode | 2. HV (Hybrid Vehicle) Mode |
|---|---|
| • A plug-in charge control system has been adopted, which allows electrical power to be supplied to the HV high voltage battery from external power source, such as an electrical socket or charger.<br>• When the HV battery is sufficiently charged, the vehicle will basically run on the power of the motor.<br>• If the vehicle exceeds 135 km/h or accelerates suddenly when traveling in EV mode, the gasoline engine and motor work together to power the vehicle. | • During light acceleration at low speeds, the vehicle is powered by the electric motor. The gasoline engine is shut off.<br>• During normal driving, the vehicle is powered mainly by the gasoline engine. The gasoline engine also powers the generator to recharge the battery assembly and to drive the motor.<br>• During full acceleration, such as climbing a hill, both the gasoline engine and the electric motor power the vehicle.<br>• During deceleration, such as when braking, the vehicle regenerates the kinetic energy from the wheels to produce electricity that recharges the battery assembly.<br>• While the vehicle is stopped, the gasoline engine and electric motor are off, however, the vehicle remains on and operational. |

# 4. Results

It is worth noting that in the ANL tests, the vehicle was tested with some defined emission tests and specific driving strategies, not necessarily reproducing the vehicle's real use, e.g., avoiding the ICE to fully charge the battery. For the training of the model, it is crucial, albeit not trivial, to select appropriate input parameters that influence the output ("target" in AI jargon). For example, while $CO_2$ emissions are directly related to fuel consumption, other pollutant emissions may be directly related to aftertreatment catalytic converter temperature. While catalytic converter temperature is not necessary as an input parameter for a digital twin predicting $CO_2$, it is crucial for predicting other pollutant emissions. See discussion in Tomanik et al., Maria et al. and Krysmon et al. [25, 26, 30]. In general, the input parameters have been classified according to the systems of the vehicle. As for other complex systems, the process of evaluating emissions from hybrid vehicles is based on two interrelated stages: data collection and subsequent analysis, see **Figures 6** and **7**. In general, data collection would involve conducting emissions measurements under various conditions, including simulations, on test benches, chassis dynamometers, and through on-board measurements in real-world operating conditions of the vehicle. The obtained data would be subject to event detection, to allow for their classification into critical and non-critical sequences, see **Figure 8**. Additionally, the data collection would include parameters that characterize the model's resilience, obtained during daily operations, including actual routes and fleet inspections. In this work, we use just a vehicle dynamometer emission test as an example.

The ANL file "62104013 Test Data" was used for this particular study. According to the ANL site, the dataset consists of 10 repetitions of the Highway emission cycle (HWFET). This specific data set consists of 78,263 lines and 318 columns. For better documentation and auxiliary plots, the txt. file was converted into an Excel file, a fragment seen in **Figure 9**. Some data cleaning and organization were done to help speed up the pre-analysis. After input in the code, it automatically filters the few lines ("instances" in AI jargon) with missing values and other filters that the user chooses to define.

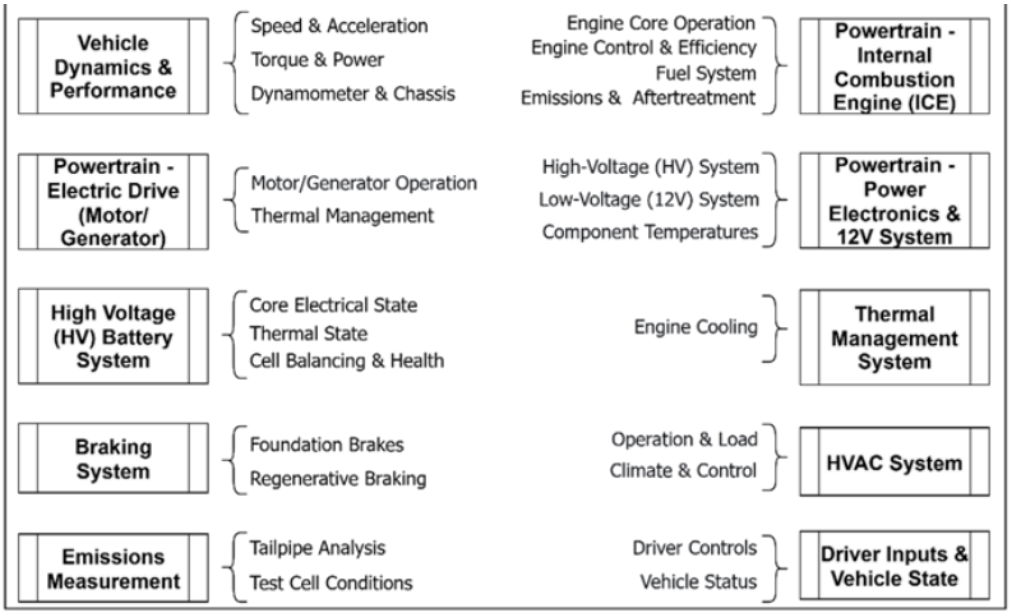


**Figure 8.** Proposed Classification of Input Parameters.

| Dyno_Spd [mph] | Dyno_TractiveForce[N] | Dyno_LoadCell[N] | Distance[mi] | Dyno_Spd_Front[mph] | Dyno_TractiveForce_Front[N] | Dyno_LoadCell_Front[N] | Dyno_Spd_Rear[mph] | Dyno_LoadCell_Rear[N] | Dyno_TractiveForce_Rear[N] | DilAir_RH[%] | Tailpipe_Press[inH2O] | Cell_Temp[C] | Cell_RH[%] | Cell_Press[inHg] | Tire_Front_Temp[C] | Drive_Trace_Schedule[mph] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.002 | -12.583 | -21.214 | 0.001 | 0.002 | -9.613 | 3.695 | 0 | -24.909 | -2.97 | 1.66 | -0.232 | 21.45 | 61.289 | 29.123 | 20.225 | 0 |
| 0.001 | -12.574 | -21.294 | 0.001 | 0.001 | -9.625 | 3.232 | 0 | -24.526 | -2.949 | 1.649 | -0.279 | 21.45 | 61.288 | 29.123 | 20.226 | 0 |
| 0.002 | -12.568 | -19.908 | 0.001 | 0.002 | -9.624 | 4.554 | 0 | -24.462 | -2.944 | 1.655 | -0.3 | 21.452 | 61.287 | 29.123 | 20.228 | 0 |
| 0.002 | -12.594 | -18.372 | 0.001 | 0.002 | -9.611 | 5.588 | 0 | -23.96 | -2.983 | 1.653 | -0.307 | 21.452 | 61.286 | 29.123 | 20.229 | 0 |
| 0.002 | -12.619 | -20.98 | 0.001 | 0.002 | -9.652 | 3.092 | 0 | -24.072 | -2.967 | 1.657 | -0.279 | 21.452 | 61.285 | 29.123 | 20.227 | 0 |
| 0.002 | -12.613 | -20.515 | 0.001 | 0.002 | -9.621 | 3.628 | 0 | -24.143 | -2.991 | 1.651 | -0.253 | 21.452 | 61.286 | 29.123 | 20.228 | 0 |
| 0.002 | -12.604 | -22.783 | 0.001 | 0.002 | -9.608 | 2.749 | 0 | -25.531 | -2.996 | 1.654 | -0.263 | 21.452 | 61.286 | 29.123 | 20.227 | 0 |
| 0.002 | -12.647 | -18.721 | 0.001 | 0.002 | -9.608 | 4.984 | 0 | -23.705 | -3.038 | 1.659 | -0.243 | 21.452 | 61.286 | 29.123 | 20.228 | 0 |
| 0.002 | -12.627 | -20.761 | 0.001 | 0.002 | -9.608 | 3.654 | 0 | -24.415 | -3.019 | 1.652 | -0.248 | 21.452 | 61.286 | 29.123 | 20.229 | 0 |
| 0.002 | -12.645 | -19.197 | 0.001 | 0.002 | -9.659 | 4.605 | 0 | -23.802 | -2.986 | 1.649 | -0.226 | 21.453 | 61.286 | 29.123 | 20.229 | 0 |
| 0.002 | -12.614 | -18.223 | 0.001 | 0.002 | -9.623 | 5.589 | 0 | -23.812 | -2.991 | 1.651 | -0.209 | 21.454 | 61.285 | 29.123 | 20.23 | 0 |
| 0.001 | -12.636 | -21.423 | 0.001 | 0.001 | -9.608 | 3.086 | 0 | -24.509 | -3.027 | 1.659 | -0.242 | 21.454 | 61.284 | 29.123 | 20.227 | 0 |
| 0.002 | -12.668 | -20.198 | 0.001 | 0.002 | -9.641 | 3.639 | 0 | -23.837 | -3.026 | 1.662 | -0.225 | 21.455 | 61.283 | 29.123 | 20.228 | 0 |
| 0.002 | -12.663 | -21.801 | 0.001 | 0.002 | -9.64 | 2.701 | 0 | -24.502 | -3.023 | 1.661 | -0.204 | 21.455 | 61.284 | 29.123 | 20.229 | 0 |
| 0.002 | -12.627 | -20.971 | 0.001 | 0.002 | -9.665 | 2.849 | 0 | -23.82 | -2.962 | 1.663 | -0.234 | 21.455 | 61.283 | 29.123 | 20.23 | 0 |
| 0.001 | -12.617 | -22.217 | 0.001 | 0.001 | -9.629 | 2.49 | 0 | -24.707 | -2.988 | 1.654 | -0.226 | 21.454 | 61.283 | 29.123 | 20.232 | 0 |
| 0.001 | -12.611 | -20.968 | 0.001 | 0.001 | -9.632 | 2.861 | 0 | -23.829 | -2.979 | 1.649 | -0.262 | 21.455 | 61.283 | 29.123 | 20.229 | 0 |
| 0.002 | -12.61 | -20.387 | 0.001 | 0.002 | -9.626 | 4.601 | 0 | -24.989 | -2.984 | 1.656 | -0.31 | 21.456 | 61.283 | 29.123 | 20.227 | 0 |
| 0.002 | -12.671 | -20.296 | 0.001 | 0.002 | -9.634 | 3.46 | 0 | -23.757 | -3.037 | 1.66 | -0.283 | 21.457 | 61.282 | 29.123 | 20.228 | 0 |
| 0.002 | -12.6 | -21.994 | 0.001 | 0.002 | -9.626 | 2.451 | 0 | -24.445 | -2.974 | 1.652 | -0.251 | 21.457 | 61.281 | 29.123 | 20.229 | 0 |

**Figure 9.** Partial View of the Dataset Used in This Work.

The process of evaluating emissions from hybrid vehicles is based on two interrelated stages: data collection and subsequent analysis (**Figure 10**). Data collection involves conducting emissions measurements under various conditions, including simulations, on test benches, chassis dynamometers, and through mobile measurements of operating conditions of the vehicle. The obtained data are subject to event detection, which allows for their classification into critical and non-critical sequences.

The reliability and accuracy of a predictive model fundamentally depend on three key operations: accurate event detection, effective data clustering, and in-depth statistical analysis. Properly separating critical data sequences from non-critical ones is essential for building an accurate model. At the same time, the application of strict statistical controls ensures the robustness of the results, and the integration of real operational data significantly improves the adaptability and stability of the model under different conditions. For this paper, we are assuming that, within certain limitations, the Highway official emission test protocol can be used to

illustrate the use of AI.

After initial data cleaning, such as removing pre-test and post-test values, the first step is to determine which input parameters actually affect the result. Many instantaneous test parameters can be excluded from the current study. This may be because they remain constant throughout the test (e.g., ambient pressure and temperature), have a negligible effect on the result, or are interdependent (e.g., fuel consumption and $CO_2$ emissions). During the first five cycles, the vehicle operated only in EV mode, and the battery was gradually depleting. On repetitions 6 to 10, the vehicle operated in the hybrid mode, but apparently, the ICE was not fully charging the battery, only keeping the SoC (State of Charge) constant. See **Figures 11–13**.

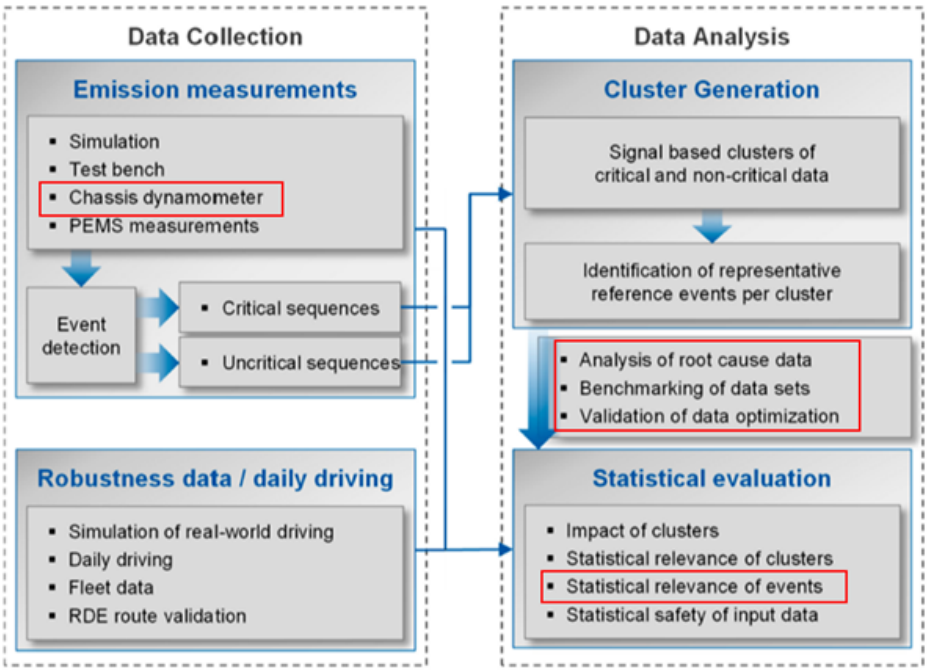


**Figure 10.** Overview of the Data Analysis Steps.

Note: Remarked in red, the ones covered in this paper.

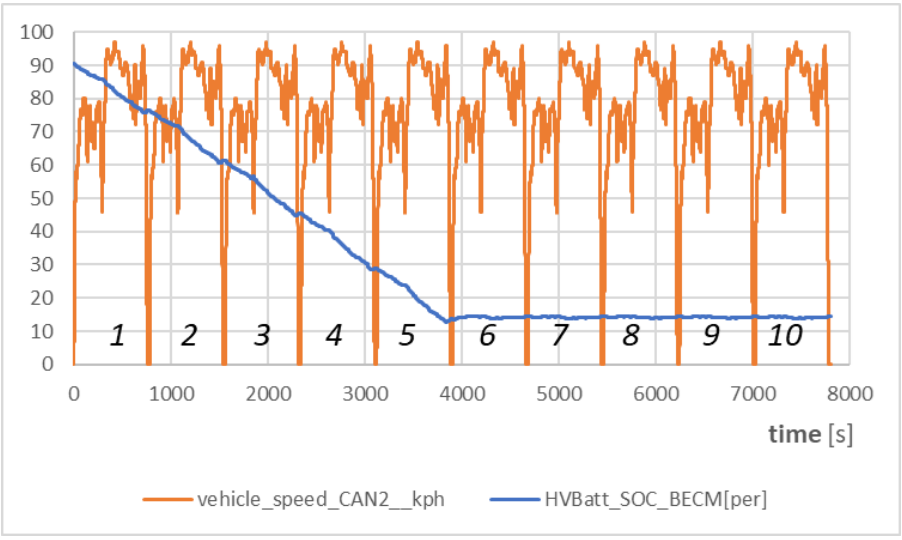


**Figure 11.** HWFET test used as case study.

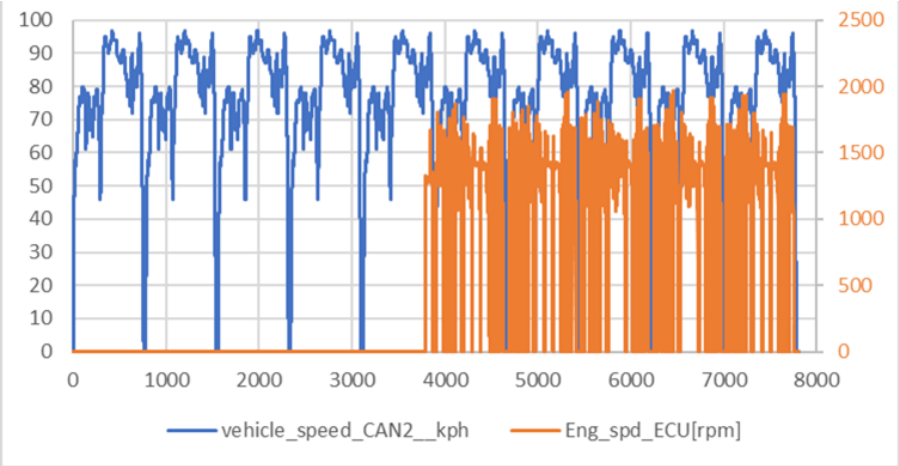


**Figure 12.** Idem of **Figure 11**, showing that the ICE starts to work only at the very end of repetition #5.

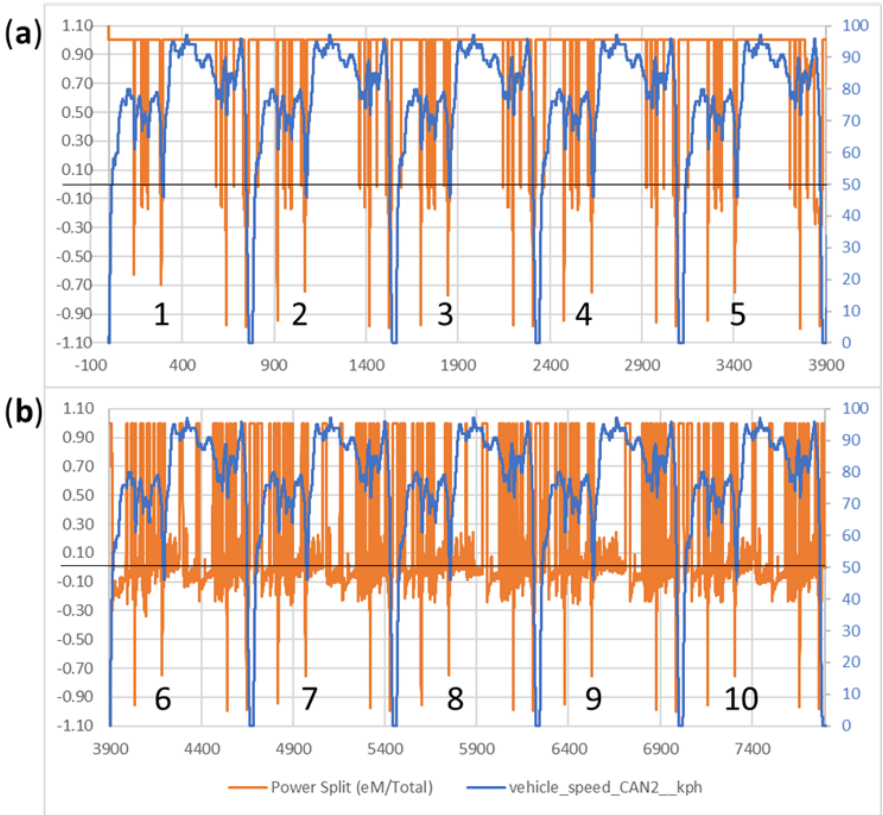


**Figure 13.** Power Share: (**a**) in the EV mode only, (**b**) in the Hybrid mode.

In addition to the original dataset parameters, two others were calculated and included in the dataset and analysis:

- **Acc**: the vehicle's instantaneous acceleration, calculated from the vehicle speed change between +0.5 and −0.5 s.
- **PS**: "Power Share", the instantaneous power ratio of the electrical motor to the total moving the vehicle (i.e., ICE + EM), calculated by:

**PS** = eM/(ICE + eM), If eM > 0 (electrical motor moving the vehicle, partially or totally);

eM/Min(eM), If eM ≤ 0 (electrical motor working as generator).

Negative values indicate when the electrical motor is working as a generator, producing power to charge the bat-

tery. **Figure 13** shows the PS over 10 repetitions. PS was 1.0 almost all the time along the first 5 repetitions, when the vehicle operated in the EV mode only, going to negative values only during vehicle decelerations, but not enough to significantly recharge the battery. Notice the SoC minor upward values in **Figure 11**.

To explore the model capabilities, three different cases were run:

- **Case A**: as input parameters, only the vehicle's instantaneous speed, its acceleration, and the battery state of charge (SoC). The first two are defined by the emission standard, the SoC is a consequence of the electrical energy consumed to move the vehicle and the regenerated power during decelerations, and by the ICE. Such a very simple approach was chosen to demonstrate the model's ability with very few input parameters, which would make it of more general use. To remark that, as mentioned before, in this specific test, the ICE was used only to keep the SoC constant, not to fully charge the battery.
- **Case B**: Instantaneous Battery Outlet coolant temperature was included as an input parameter, trying to include instantaneous information since the SoC is a cumulative value.
- **Case C**: ICE oil temperature was included as an input parameter to indicate when the engine is cold, not operating. Other ICE parameters could, of course, be used.

**Table 6** summarizes the results, showing the data Spearman and Pearson correlation, the model $R^2$ and MSE values, as well as the SHAPLEY values. **Table 6** describes the input parameters (dataset columns used for each case). Using as input parameters only Vehicle speed, Acceleration, and SoC already allowed the digital twin to achieve an $R^2$ of 0.80.

**Table 6.** Correlation Parameters and Model PS Results for the Different Cases.

| Input Parameter | Case A | | | | | Case B | | | | | Case C | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Spe** | **Pea** | **Shap** | **$R^2$** | **MSE** | **Spe** | **Pea** | **Shap** | **$R^2$** | **MSE** | **Spe** | **Pea** | **Shap** | **$R^2$** | **MSE** |
| km/h | 0.07 | 0.02 | 0.10 | | | 0.07 | 0.02 | 0.10 | | | 0.07 | 0.02 | 0.09 | | |
| Acc | 0.21 | 0.29 | 0.12 | | | 0.21 | 0.29 | 0.12 | | | 0.21 | 0.29 | 0.10 | | |
| Toil | | | | 0.80 | 0.06 | | | | 0.84 | 0.05 | 0.51 | 0.58 | 0.30 | 0.97 | 0.01 |
| T Bat_out | | | | | | 0.52 | 0.55 | 0.08 | | | 0.52 | 0.55 | 0.02 | | |
| SoC | 0.55 | 0.50 | 0.32 | | | 0.55 | 0.50 | 0.28 | | | 0.55 | 0.50 | 0.02 | | |

The inclusion of Battery temperature did not significantly increase the model accuracy, and the Battery Temperature Shapley value reflects it, showing a value lower than Vehicle speed and acceleration. Inclusion of ICE oil temperature increased the model $R^2$ to 0.97 and curiously changed the ranking of the other input parameters. The SoC that was the most influential in Cases A and B had almost no impact on the results. See discussion ahead.

# 5. Discussion

$R^2$ and MSE are common and useful model accuracy indicators, but should be seen with care. **Figure 14** shows, for the three cases from top to bottom: the PS actual value (in blue) and the model errors (in red), the plot of actual versus model, and the histograms of the errors. It can be seen that:

i) The errors are concentrated when PS is close to zero,

ii) The model was more accurate when the PS is negative, i.e., the electrical motor is working as a generator,

iii) The errors when PS is equal to 1.0 (the vertical line on the right side of the correlation plots) are indeed few. Instances when the model error was higher than 0.9 were 0.38, 0.27, and 0.08%, respectively, for Cases A, B, and C.

**Figure 15** shows the Shapley values. The bar plot shows from the influence ranking and value of each input parameter in the output but when an information of ICE use, T_oil, was included, SoC influence becomes almost negligible, not because SoC does not affect PS but because in this AI model, which of course it does, but such information was not relevant for the predictions in this dataset, where the repetitions 1 to 5 ran exclusively with eMotor while in repetitions 6 to 10, ICE was used to move the vehicle but not to fully recharge the battery. Such peculiarity illustrates the risks of using singular datasets and the need for a more comprehensive and also physical based approach done by Human experts.

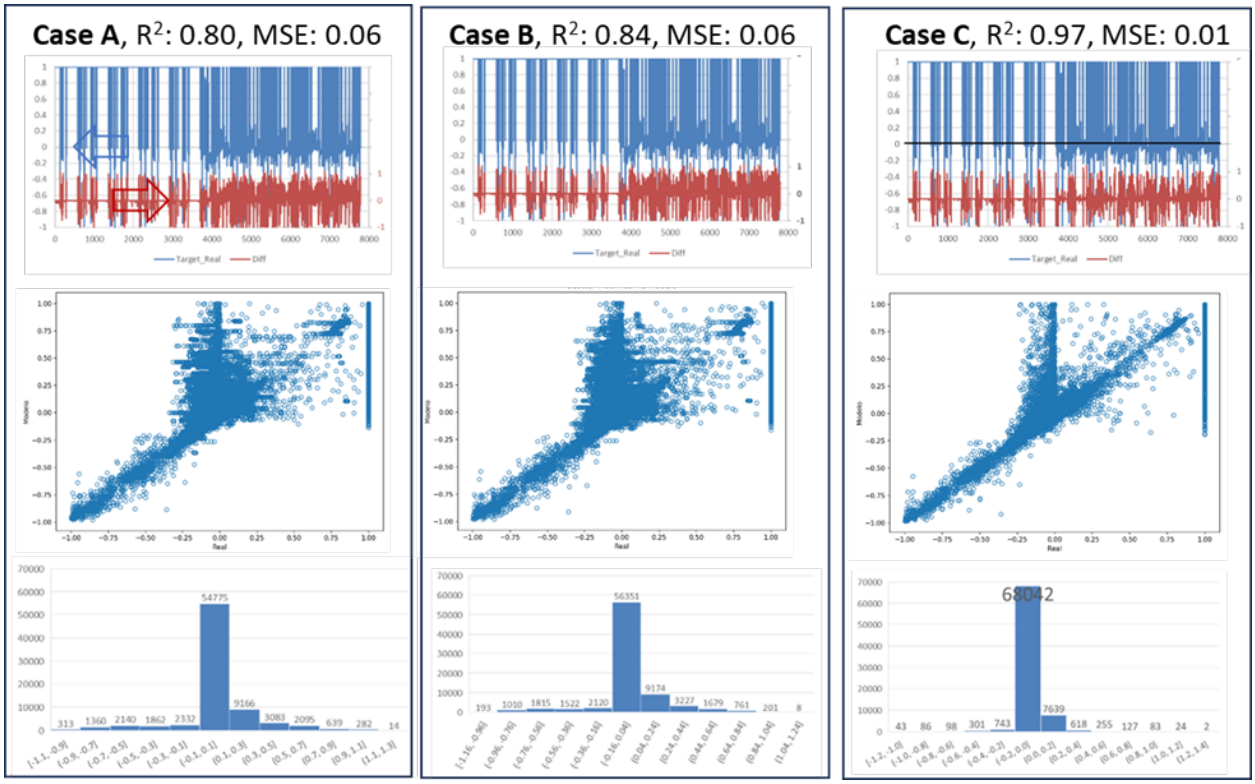


**Figure 14.** Model Results.

Note: From top to bottom: PS actual values (in blue) and the model error (in red), PS actual versus model, and the error histogram.

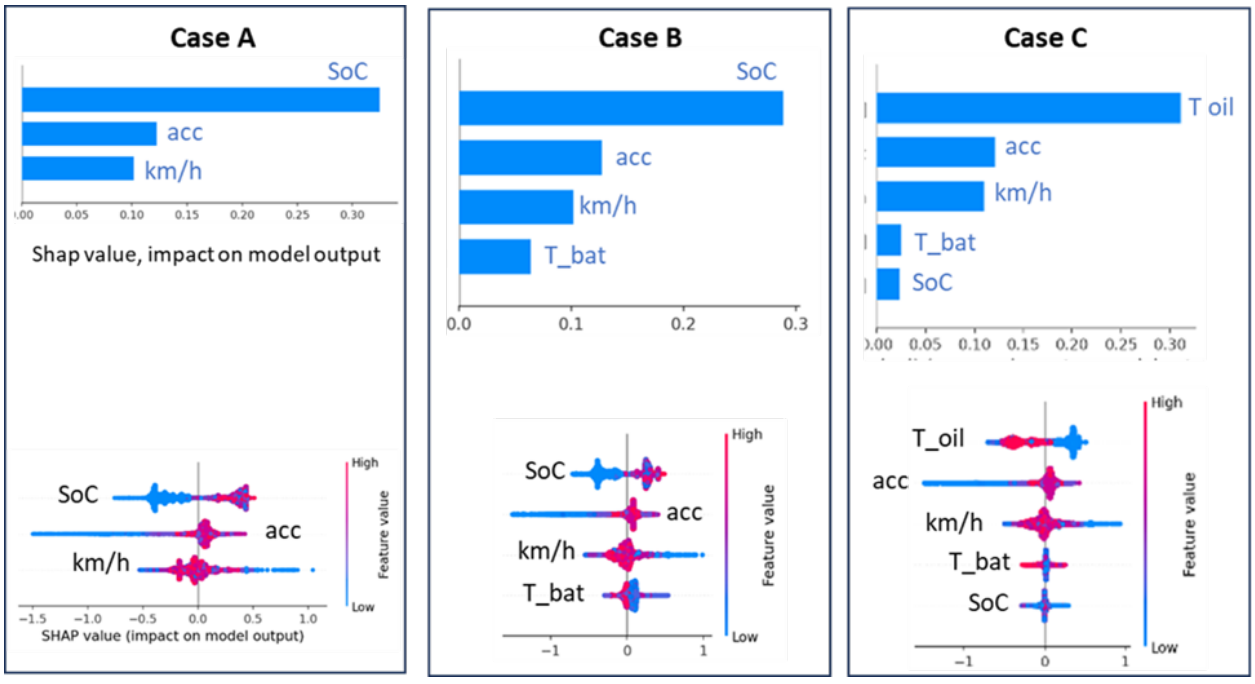


**Figure 15.** Shapley Plots.

Note: Top: input parameter ranking. Bottom: impact on the model output (PS).

The plot of impact on model output brings more information. In the plot, each dot represents one observation (in our case, the reading every 0.1).

On the X-axis, SHAP value means the impact on the

output prediction:

- Right (positive SHAP): increases predicted PS;
- Left (negative SHAP): decreases predicted PS;
- Color: Red: high Power Split (more e-motor share), Blue = low PS (more ICE share).

So, if you see red dots mostly on the right, the model "believes" that the input parameter increases PS. There are several reasons this can occur, not necessarily a physical base influence, but a data or feature interaction effect:

i) Indirect Correlation/Context Effects: High e-motor usage might occur during high acceleration, where ICE also did on repetitions 6 to 10, but not on 1 to 5.

ii) Feature Collinearity: A given parameter might be correlated with variables that actually influence the PS, in our example, T_oil, which increases with ICE use.

iii) Specific Test Conditions: In the case study, eMotor was the only responsible for moving the vehicle, and SoC was decreasing according to.

Care should be taken to avoid using correlated parameters. Model accuracy was improved with the inclusion of oil temperature, Case C. Shapley values ranking was significantly changed, with SoC dropping from first to last. A more careful analysis of the data would show that Cases A and B were modelled without any information about the ICE, and that battery SoC (and Oil temperature) had a very different behaviour on the repetitions 1 to 5 than on repetitions 6 to 10, which may explain its strong influence in the model. **Table 7** shows the model results using the ICE torque as an input parameter. Model accuracy was quite decent with an $R^2$ of 0.94 and an MSE of 0.13. The SoC ranking was very low. Also, it is worth noting that both Oil temperature and SoC are cumulative, with relatively slow changes, and are unable to be impacted by fast transients as the ones observed in PS.

To test model robustness, the model was trained with a given repetition and asked to predict another. An important parameter on ICE, and hence Hybrid vehicles, is the $CO_2$ emitted during the test cycle. The model was used to predict the instantaneous $CO_2$, having km/h, acc, PS, and Pedal accel as input parameters. **Table 8** summarizes the results, and **Figure 16** plots the measured versus the predicted values for each case.

To further investigate model generalization and robustness, the model was used to predict instantaneous $CO_2$ along the complete dataset, the 10 repetitions, after being trained with a different test cycle, the urban-driven UDDS. The UDDS cycle is composed of two sub-cycles with 13.695 s with an average speed of 31.5 km/h and several start-stops, typical of urban driving, see **Figure 17**.

Based on the previous cases, the AI model used as input parameters: km/h, Acc. PS and torque to predict the instantaneous $CO_2$. Instances with $CO_2$ lower than 10 mg/s were removed from the datasets (see discussion in **Appendix A**). The AI model accuracy was excellent: $R^2 = 0.98$, MSE = 275, and cycle accumulated error of −0.1%, see **Figure 18a**. Using this model, trained with the UDDS cycle, to predict the HWY also produced quite reasonable predictions with $R^2 = 0.88$, MSE = 598, and cycle accumulated error = 2.0%. see **Figure 18b**.

**Table 7.** Correlation Parameters and Model PS Results for the Case C including ICE torque.

| Input Parameter | Case C | | | | | Using ICE Torque | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Spe | Pea | Shap | $R^2$ | MSE | Spe | Pea | Shap | $R^2$ | MSE |
| km/h | 0.07 | 0.02 | 0.09 | 0.97 | 0.05 | 0.07 | 0.01 | 0.05 | 0.94 | 0.13 |
| Acc | 0.21 | 0.29 | 0.10 | | | 0.21 | 0.29 | 0.15 | | |
| Toil | 0.51 | 0.58 | 0.30 | | | - | - | - | | |
| T Bat_out | 0.52 | 0.55 | 0.02 | | | - | - | - | | |
| SoC | 0.55 | 0.50 | 0.02 | | | 0.55 | 0.50 | 0.03 | | |
| ICE Torque | - | - | - | | | 0.67 | 0.71 | 0.37 | | |

**Table 8.** Results when the Model* was Trained with a Given Case and Used to Predict a Different One.

| Trained with | $R^2$ | MSE | $CO_2$ Acc. Error | Predicted | $R^2$ | MSE | $CO_2$ Acc. Error |
|---|---|---|---|---|---|---|---|
| **rep 6** | 0.99 | 168 | −0.10% | **rep 10** | 0.95 | 398 | 3.45% |
| **rep 10** | 0.99 | 168 | −0.06% | **rep 6** | 0.93 | 460 | 3.1% |

Note: * km/h, Acc, PS, and Pedal accel as input parameters.

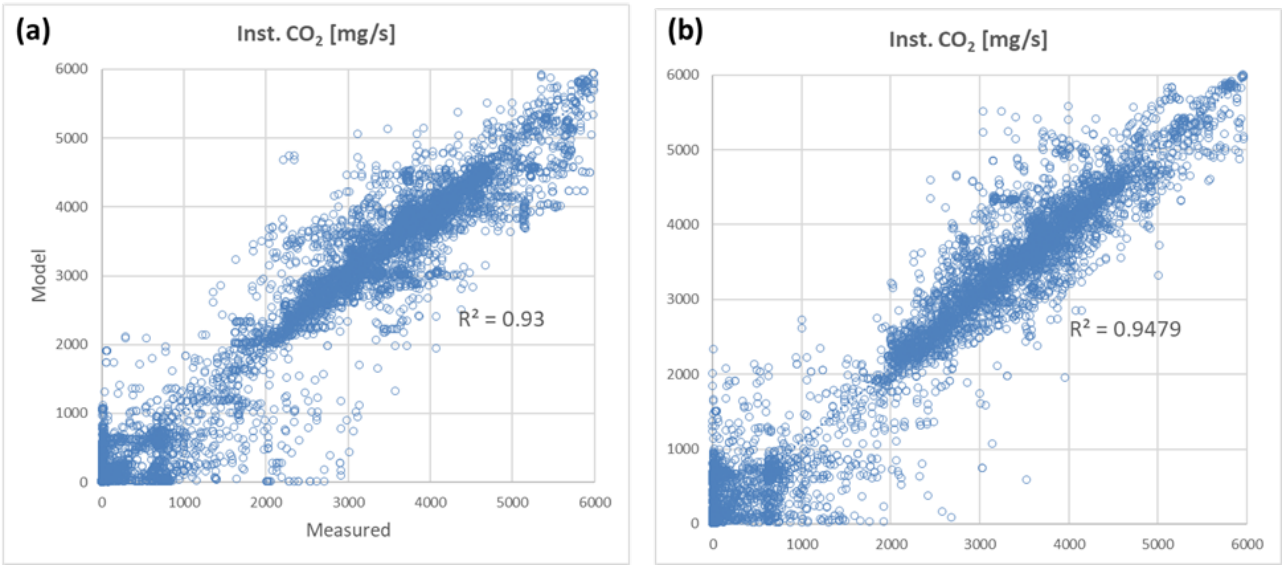


**Figure 16.** Instantaneous $CO_2$, (**a**) trained with HWY rep 10 and used to predict rep 6; (**b**) Vice-versa.

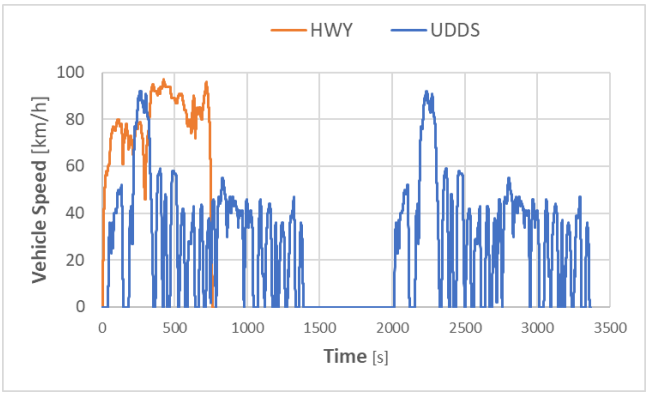


**Figure 17.** HWY and UDDS emission test cycles.

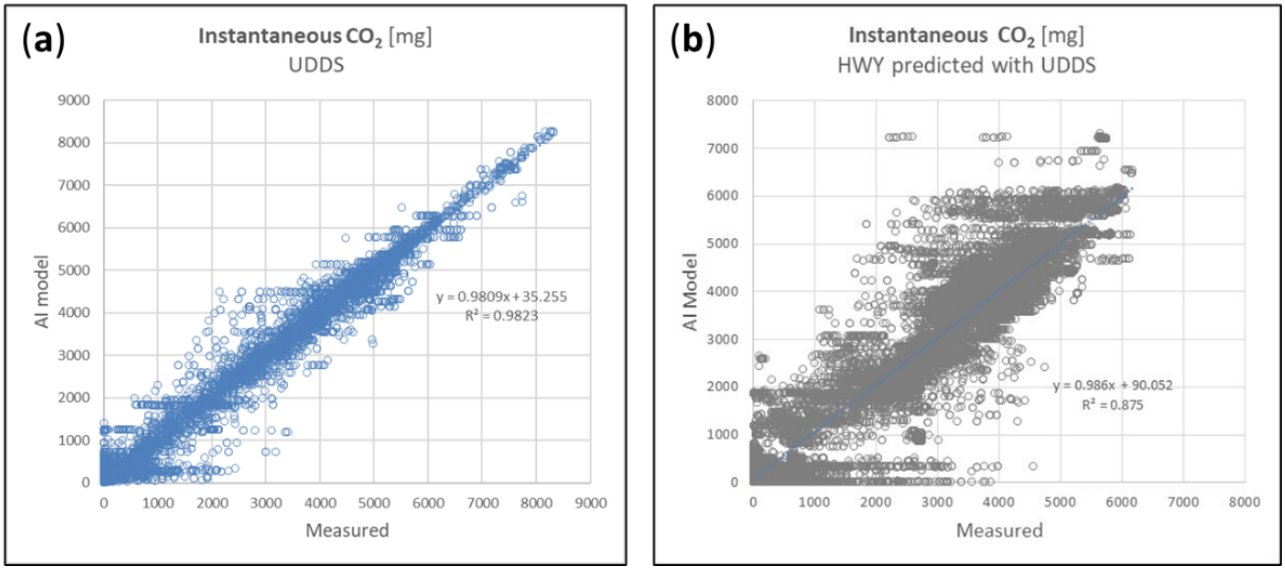


**Figure 18.** Instantaneous $CO_2$, (**a**) UDDS Cycle; (**b**) HWY cycle using the model trained with UDDS.

# 6. Conclusions

AI, and in particular, machine learning digital twins, are becoming more and more a powerful engineering tool. Its relatively low cost and quick use make more and more engineers make use of it with apparently excellent accuracy. The benefits also bring the risk of a lack of robustness and apparently correct models, but lacking physical correlation with the actual system. The model could reproduce with accuracy the dataset used for training, but may fail to fully represent the real system.

This work proposes a template for multimetric evaluation of digital twins in emissions testing of hybrid vehicles and illustrates it using the RAV4 Argonne dataset. Within this template, the implementation of standardized procedures for data preprocessing and cleansing increases the quality and correctness of input information, thereby minimizing the risks of errors and incorrect conclusions. Furthermore, the standardization of the application of Explainable Artificial Intelligence (XAI) methods allows for the formalization of the process of interpreting model results, significantly enhancing the level of transparency and trust in machine learning systems. Thus, the standardization of methods for utilizing multimetric processing templates not only ensures high quality and reliability of research but also formalizes the interpretation of model results, making its behavior more transparent for domain specialists, enhancing trust, and promoting consistency in assessments prior to the application of machine learning systems in critical scenarios while adhering to ethical and regulatory requirements.

The multimetric evaluation template for digital twins in critical domains should encompass:

- A comprehensive multi-metric assessment of performance, including reliability indicators.
- Quantitative evaluation of explainability and uncertainty for transparency.
- Rigorous quality control of data and assessment of bias/fairness.
- Adaptability options for heterogeneous or decentralized datasets.
- Standardized reporting and reproducibility protocols.
- Ethical and socio-technical aspects of impact.

This research demonstrates that high predictive accuracy is insufficient to guarantee reliability: even a seemingly well-performing model may inadequately represent physically relevant mechanisms if the training data and the selection of input variables are not aligned with domain knowledge. This underscores the necessity of integrating human expert judgments and physical considerations in the evaluation of machine learning (ML), particularly when digital twins are employed for the interpretation or optimization of critical systems, such as hybrid powertrains.

In a broader sense, the proposed template and case study of the hybrid vehicle represent a practical step towards more standardized evaluation practices in critical applications of machine learning. Future work will extend this framework to real-world road measurements and PEMS data, encompass a wider range of operational and environmental conditions, and assess how the same multimetric framework can be adapted to other critical domains while maintaining transparency and traceability of model behavior.

# Author Contributions

Conceptualization, N.M. and E.T.; methodology, all; software, E.R., validation, E.T., F.F.R.; formal analysis, N.M.; investigation, E.T.; writing—original draft preparation, N.M.; writing—review and editing, N.M. and E.T. All authors have read and agreed to the published version of the manuscript.

# Funding

This work received no external funding.

# Institutional Review Board Statement

Not applicable.

# Informed Consent Statement

Not applicable.

# Data Availability Statement

All data used in this work were downloaded from Ar-

gonne National Laboratory, Downloadable Dynamometer Database (https://www.anl.gov/es/downloadable-dynamometer-database). Additional data details, as used in this work, are available after a reasonable request.

## Acknowledgments

The authors acknowledge the support on analyzing the data and the valuable discussions of Leonardo Sutti. We would like to be grateful to the Argonne National Laboratory in sharing such valuable and detailed test data.

## Conflicts of Interest

The authors declare no conflict of interest.

## Appendix A. Workflow of a Machine Learning Tool for Vehicle Emission Tests

**Table A1** shows a modelling workflow based on the one proposed by Jiang[17] with adaptations, and examples for vehicle test emissions.

**Table A1.** Recommended Procedure to Create and Use the AI Model for emission Tests.

| | Example | Recommendations |
|---|---|---|
| **1. Raw Data** | | |
| Data types | The data used in the current and in the author's previous works[20, 28, 29] came from different sources with different parameters, frequency, and nomenclature for the same parameter. | Format the data to an uniform one to avoid having to declare it inside the model. |
| **2. Data Processing** | | |
| Missing or invalid values | The test data contains pre- and post-test data<br>Small $CO_2$ and pollutant emissions values when the vehicle is running only with the electrical motor. | Use data visualization tools to check data consistency.<br>Do filtering when necessary. See **Figure A1**. |
| Denoising | The first 5 repetitions contain small, not valid, values of $CO_2$ that "contaminated" the machine learning, which, although not affecting the total cycle value | Using only the last 5 repetitions, when ICE was run, to predict $CO_2$.<br>Apply Filters. See **Figure A1**. |
| Feature extraction | Identification of critical parameters. See discussion of the inclusion of SoC and Oil temperature. | Use a human expert. Avoid using correlated parameters that may deteriorate model generalization. |
| Dimensionality Reduction | Several test data parameters, such as the temperature chamber, tyre pressure, etc., are monitored just to attend the homologation procedure. | Remove the obvious useless, and constant parameters.<br>Use PCA (Principal Component Analysis) tools. |
| Feature Normalization & Standardization | In the test data, while PS goes from −1 to +1, $CO_2$ goes from 0 to 9000. The used Random Forest Model works well with such different ranges, but, for example, ANN does not. | Normalize data for ANN.<br>Consider returning to the absolute values when analyzing to "return to the actual case" |
| Data split (train/validation) | 70/30% is the more common ratio, but notice that some cycles, like the UDDS, include long periods of "non-test" that can contaminate the model's robustness. | |
| Data Augmentation | Some relevant parameters, such as the PS, can appear splitted in different parameters in the dataset. Others, like vehicle acceleration, can be easily derived from others. | If necessary, calculate and add important parameters not included in the original dataset. |
| Confidence, absence of undesirable bias of data source | In the case study of 10 cycle repetitions, the ECU was programmed to run only as electrical in the first 5 and as hybrid, but not fully charging the battery, in the last 5 repetitions. | Pay attention to dataset features that can deteriorate model generalization.<br>Use of different tests, for example, combining urban and highway cycles to check model robustness. |

**Table A1.** *Cont.*

| | Example | Recommendations |
|---|---|---|
| **3. Model Evaluation** | | |
| Select metrics critical to the study objective | | Use accumulated, instead of instantaneous, emission values for emissions. |
| Critical analysis of relevant parameters | | Use model Shapley values. |
| Check model generalization | | Use the model to predict datasets different from the one used for training. |

As mentioned, the dataset contains 10 repetitions of the Highway emission test cycle, with the first five repetitions the vehicle being moved only by the electrical motor. However, the measurement equipment read very small quantities of $CO_2$, with no influence on the accumulated value, but that degrades the model accuracy if used to train the model. **Figure A1** shows the accumulated $CO_2$ and the model error with different filters for the $CO_2$ values: No filtering (NF), higher than 1 (F1), and higher than 10 mg/s. The accumulated measured $CO_2$ along the cycle is little changed with the filters: 9609, 9612, and 9606 g, respectively, for NF, F1, and F10. However, model accuracy was significantly better with filtering. The cycle accumulated error decreased from 11% with no filter to 1.0% when instances with $CO_2$ lower than 10% were eliminated, see **Figure A1**.

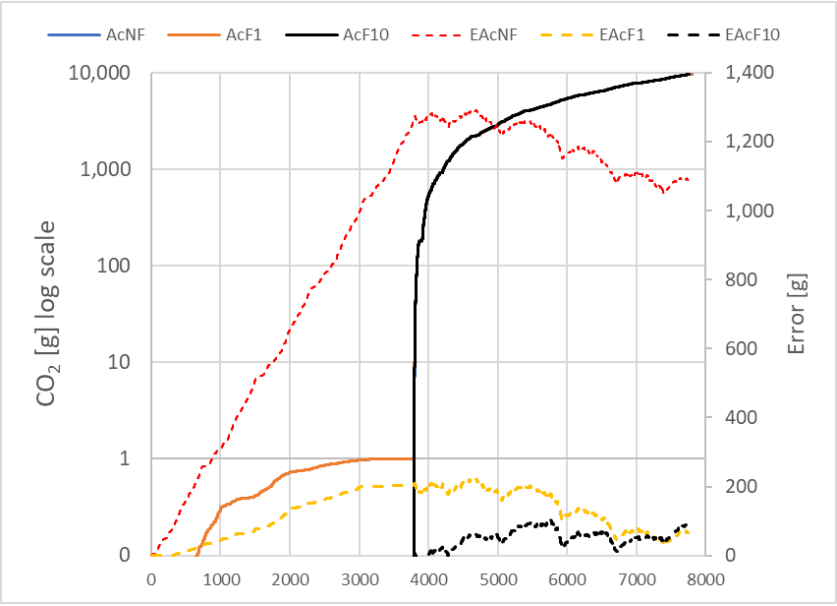


**Figure A1.** Accumulated measured $CO_2$ and model errors when different filters for instantaneous $CO_2$ is applied.

Note: No Filter (NF), higher than 1 mg/s (F1), and higher than 10 mg/s (F10).

# References

[1] International Organization for Standardization, & International Electrotechnical Commission, 2022. ISO/IEC 23053:2022—Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML). ISO: Geneva, Switzerland. Available from: https://www.iso.org/standard/74438.html

[2] Tabassi, E., 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0) (No. NIST AI 100-1). National Institute of Standards and Technology (U.S.): Gaithersburg, MD, USA. DOI: https://doi.org/10.6028/NIST.AI.100-1

[3] Raja, V.J., M, D., Solaimalai, G., et al., 2024. Machine Learning Revolutionizing Performance Evaluation: Recent Developments and Breakthroughs. In Proceedings of the 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS),

Coimbatore, India, 10 July 2024; pp. 780–785. DOI: https://doi.org/10.1109/ICSCSS60660.2024.10625103

[4] Anžel, A., Heider, D., Hattab, G., 2023. Interactive polar diagrams for model comparison. Computer Methods and Programs in Biomedicine. 242, 107843. DOI: https://doi.org/10.1016/j.cmpb.2023.107843

[5] Carvalho, J.B.S., Rodriguez, V.J., Torcinovich, A., et al., 2025. Rethinking Robustness in Machine Learning: A Posterior Agreement Approach. arXiv preprint. arXiv:2503.16271. DOI: https://doi.org/10.48550/ARXIV.2503.16271

[6] Hassija, V., Chamola, V., Mahapatra, A., et al., 2024. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation. 16(1), 45–74. DOI: https://doi.org/10.1007/s12559-023-10179-8

[7] Sewada, R., Jangid, A., Kumar, P., et al., 2023. Explainable Artificial Intelligence (XAI). Journal of Nonlinear Analysis and Optimization. 13(1), 41–47. DOI: https://doi.org/10.36893/JNAO.2022.V13I02.041-047

[8] Mothilal, R.K., Sharma, A., Tan, C., 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27 January 2020; pp. 607–617. DOI: https://doi.org/10.1145/3351095.3372850

[9] Hanselle, J., Heid, S., Fürnkranz, J., et al., 2024. Probabilistic Scoring Lists for Interpretable Machine Learning. arXiv preprint. arXiv: 2407.21535. DOI: https://doi.org/10.48550/arXiv.2407.21535

[10] Devarasetty, N., 2024. Optimizing data engineering for AI: improving data quality and preparation for machine learning application. Research and Analysis Journal. 7(3), 1–29. DOI: https://doi.org/10.18535/raj.v7i03.397

[11] Kamrud, A., Borghetti, B., Schubert Kabban, C., 2021. The Effects of Individual Differences, Non-Stationarity, and the Importance of Data Partitioning Decisions for Training and Testing of EEG Cross-Participant Models. Sensors. 21(9), 3225. DOI: https://doi.org/10.3390/s21093225

[12] Ferrara, E., 2023. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci. 6(1), 3. DOI: https://doi.org/10.3390/sci6010003

[13] Cloots, A.S., 2019. Blockchain and the Law: The Rule of Code. By Primavera De Filippi and Aaron Wright. The Cambridge Law Journal. 78(1), 213–217. DOI: https://doi.org/10.1017/S0008197319000084

[14] Apicella, A., Isgrò, F., Prevete, R., 2025. Don't push the button! Exploring data leakage risks in machine learning and transfer learning. Artificial Intelligence Review. 58(11), 339. DOI: https://doi.org/10.1007/s10462-025-11326-3

[15] Wolff, R.F., Moons, K.G.M., Riley, R.D., et al., 2019. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Annals of Internal Medicine. 170(1), 51–58. DOI: https://doi.org/10.7326/M18-1376

[16] Hu, H., Liu, M., Yuan, D., et al., 2017. A block based encoding approach for improving sliding window network coding in wireless networks. In Proceedings of the 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, December 2017; pp. 300–304. DOI: https://doi.org/10.1109/CompComm.2017.8322560

[17] Jiang, W., 2021. Applications of deep learning in stock market prediction: Recent progress. Expert Systems with Applications. 184, 115537. DOI: https://doi.org/10.1016/j.eswa.2021.115537

[18] Khiari, J., Olaverri-Monreal, C., 2023. Uncertainty-Aware Vehicle Energy Efficiency Prediction Using an Ensemble of Neural Networks. IEEE Intelligent Transportation Systems Magazine. 15(5), 109–119. DOI: https://doi.org/10.1109/MITS.2023.3268032

[19] Tomanik, E., Jimenez-Reyes, A.J., Tomanik, V., et al., 2023. Machine-Learning-Based Digital Twins for Transient Vehicle Cycles and Their Potential for Predicting Fuel Consumption. Vehicles. 5(2), 583–604. DOI: https://doi.org/10.3390/vehicles5020032

[20] Ganesh, A.H., Xu, B., 2022. A review of reinforcement learning based energy management systems for electrified powertrains: Progress, challenge, and potential solution. Renewable and Sustainable Energy Reviews. 154, 111833. DOI: https://doi.org/10.1016/j.rser.2021.111833

[21] SAE International Technical Standard, 2023. Recommended Practice for Measuring the Exhaust Emissions and Fuel Economy of Hybrid-Electric Vehicles, Including Plug-in Hybrid Vehicles. SAE International Technical Standard: Warrendale, PA, USA. DOI: https://doi.org/10.4271/J1711_202302

[22] European Union, 2018. Regulations: Commission Regulation (EU) 2018/1832. Official Journal of the European Union. L 301/1. Available from: https://www.stradalex.eu/en/se_src_publ_leg_eur_jo/toc/leg_eur_jo_3_20181127_301/doc/ojeu_2018.301.01.0001.01

[23] International Council on Clean Transportation (ICCT), 2017. Real-driving Emissions Test Procedure for Exhaust Gas Pollutant Emissions of Cars and Light Commercial Vehicles in Europe. International Council on Clean Transportation (ICCT): Washington, DC, USA. Available from: https://theicct.org/publication/real-driving-emissions-test-procedure-for-exhaust-gas-pollutant-emissions-of-cars-and-light-commercial-vehicles-in-europe/?gad_source=1&gad_campaignid=22639629046&gbraid=0AAAAA_pFlefpgv0teP7NMaKgMc-h4PbHr&gclid=CjwKCAiA-__MBhAKEiwASBmsBIdx60BAmW1OXrawYvufoH2OjTNdDBaHQEReNGxhz7uWr0WwjWzlFhoCmXwQAvD_B

wE

[24] Scikit-learn developers, n.d. RandomForestRegressor. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (cited 2 January 2025).

[25] Tomanik, E., Miedviedieva, N., Tomanik, V., et al., 2025. Use of Digital Twins to Analyze and Predict CO2 and Emissions on Hybrid Vehicles. In: Slavinska, O., Danchuk, V., Kunytska, O., et al. (Eds.). Intelligent Transport Systems: Ecology, Safety, Quality, Comfort, Lecture Notes in Networks and Systems. Springer Nature: Cham, Switzerland. pp. 135–147. DOI: https://doi.org/10.1007/978-3-031-87379-9_13

[26] Maria, T.M., Tomanik, E., Ivanytska, A., et al., 2025. Engine Emissions Test Analysis Model Based on Instantaneous OBD Reading and AI. In: Slavinska, O., Danchuk, V., Kunytska, O., et al. (Eds.). Intelligent Transport Systems: Ecology, Safety, Quality, Comfort, Lecture Notes in Networks and Systems. Springer Nature: Cham, Switzerland. pp. 161–171. DOI: https://doi.org/10.1007/978-3-031-87379-9_15

[27] Lundberg, S.M., Lee, S.I., 2017. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems. Curran Associates, Inc.: Red Hook, NY, USA. Available from: https://www.bibsonomy.org/bibtex/172c147d69a1d43d7b64860b04a3295c6

[28] Argonne National Laboratory, n.d. Downloadable Dynamometer Database. Available from: https://www.anl.gov/es/downloadable-dynamometer-database (cited 3 February 2025).

[29] Toyota Motor Corporation, 2020. RAV4 PRIME RAV4 Plug-in Hybrid—Gasoline-Electric Hybrid Synergy Drive—Hybrid Vehicle Dismantling Manual. Toyota Motor Corporation: Toyota, Japan. Available from: https://content.instructables.com/FSJ/JHT2/LDREAKBB/FSJJHT2LDREAKBB.pdf

[30] Krysmon, S., Claßen, J., Pischinger, S., et al., 2023. RDE Calibration—Evaluating Fundamentals of Clustering Approaches to Support the Calibration Process. Vehicles. 5(2), 404–423. DOI: https://doi.org/10.3390/vehicles5020023