future transportation

MDPI

Article

A Comparative Evaluation of Machine-Learning Models for Road Surface Roughness Forecasting in ITSs

Riccardo Ceriani *, Leonardo Cameli, Margherita Pazzini, Valeria Vignali and Claudio Lantieri

Department of Civil, Environmental and Material (DICAM) Engineering, University of Bologna, 40136 Bologna, Italy

* Correspondence: riccardo.ceriani3@unibo.it

Abstract

The forecasting of road surface conditions is a pivotal component for intelligent transportation systems, in terms of supporting maintenance planning, safety and mobility management. The increasing availability of large-scale monitoring data, collected from passenger vehicle fleets, enables the development of data-driven forecasting approaches. However, systematic comparisons between classical time-series models and machine-learning methods in this context remain limited. The proposed benchmarking framework evaluates direct road surface roughness forecasts at 1-, 7-, 14-, 30-, and 90-day horizons using multi-year vehicle-derived data collected across heterogeneous road segments. Daily roughness indicators are derived from raw measurements and modeled following a consistent, segment-wise experimental protocol. The proposed analysis involves the evaluation of multiple machine-learning regressors including Ridge, Random Forest and Gradient Boosting which are trained on lagged observations and rolling statistics. Performance of the models is assessed using two error metrics: unweighted and uncertainty-aware weighted. Findings indicate significant variations in predictive accuracy and robustness across models and segments, emphasizing the influence of feature-based learning strategies and data-quality weighting. The research provides a scalable and transparent methodology for evaluating forecasting models on vehicle-based road monitoring data, contributing practical guidance for the deployment of artificial intelligence in Intelligent Transport Systems (ITSs).

Keywords: road maintenance; machine learning; intelligent transport systems

Academic Editors: Zhe Dai, Chao Gao, Ang Ji and Shi Dong

Received: 27 April 2026
Revised: 17 June 2026
Accepted: 23 June 2026
Published: 26 June 2026

1. Introduction

The surface quality of road networks is a paramount important factor and key determinant of transport system performances [1]. Therefore, its assessment is a core concern in pavement management systems and Intelligent Transportation Systems (ITSs) [2,3]. The presence of surface irregularities (e.g., cracks, potholes, roughness, etc.) mainly affects driver safety [4–6] and generally reduces reliability [7] since rougher or deteriorated pavements can influence vehicle stability [8], braking behavior [9], and crash occurrence or severity [10].

Furthermore, it exhibits a broad set of operational, economic, environmental, and driver comfort outcomes [11–13]. Poor pavement condition is associated with higher vehicle dynamic loads [14], reduced driving comfort [15] and increased operating costs [16].

Due to these wide-ranging effects, continuous road condition monitoring is essential for effective infrastructure asset management. At the network level, agencies typically track a combination of functional indicators, including roughness, rutting, cracking, surface distress, skid resistance and bearing capacity in order to support maintenance prioritization and resource allocation [17–19]. Among functional indicators, road roughness is one of the most widely adopted [20] because it directly reflects pavement serviceability and can be measured comparatively efficiently over large networks. In addition, it is widely utilized by road agencies to select the treatment type of pavement sections [21,22]. The International Roughness Index (IRI) is an established longitudinal roughness metric, which has become the global reference standard for quantifying longitudinal unevenness [23]. IRI is a dimensionless indicator (mm/m or km/m) computed from a measured longitudinal profile through a quarter-car model [24,25]. It expresses the accumulated suspension motion of a reference vehicle per unit distance traveled. Due to the fact that IRI is characterized by standardization, interpretability, and a robust correlation with ride quality, it is employed extensively in pavement management systems and condition surveys on a global scale.

Despite its central role, common pavement monitoring remains constrained by important operational limitations. Conventional survey methodologies are predicated on the utilization of specialized inspection vehicles (i.e., inertial profilers, laser-based systems, dedicated field campaigns, etc.) that are costly, logistically demanding, and often necessitate manual data collection [26]. Especially on secondary and lower-volume roads, full-network inspections cannot usually be performed at the frequency because the extent of these networks and the cost of conventional surveys limit continuous monitoring [27]. Historically, pavement-condition data have often been collected at annual or multi-annual intervals, with some agencies also relying on partial or sample-based surveys to reduce costs. While these practices can provide useful insights of network condition, their temporal resolution is generally insufficient for time-sensitive maintenance decisions and for tracking short-term changes in pavement performance. As a result, early deterioration trends, seasonal variability and localized abnormal conditions may remain undetected until they become more severe [28,29].

These shortcomings are particularly significant in the context of future ITSs, where infrastructure monitoring is increasingly moving toward continuous, networked, and scalable sensing architectures [30]. The emergence of connected vehicles, embedded sensors, smartphones, and fleet-based telematics is transforming the way to monitor road conditions [31–33]. This transition enables a move from episodic surveying to quasi-continuous monitoring and opens the possibility of real-time infrastructure awareness. Recent work on connected-vehicle and floating car data (FCD) shows that vehicle-derived observations can support network-level estimation of pavement roughness and related indicators at much finer spatial and temporal granularity than traditional surveys [34–36].

Forecasts of future surface roughness can support preventive maintenance planning and budget allocation by identifying intervention needs [37]. Indeed, pavement preservation actions can reduce user costs and environmental impacts by maintaining smoother surfaces and avoiding more severe deterioration [38]. In this respect, connected vehicles and FCD data are particularly promising because they provide repeated observations over time on the same segments, thereby creating the conditions for time-series modeling and data-driven forecasting [39,40]. As connected fleets expand and multi-year archives accumulate, segment-level condition trajectories can be analyzed not only to infer the present state of the network but also to anticipate future deterioration patterns. This shift from descriptive monitoring to predictive intelligence is fully consistent with the broader evolution of ITSs toward proactive and adaptive infrastructure management.

The present study contributes to the literature by proposing a unified benchmarking framework for daily road surface roughness forecasting based on multi-year vehicle-derived observations collected over heterogeneous road segments. The study compares multiple machine-learning regressors trained on lagged values and rolling statistics. By adopting a consistent segment-wise experimental design and evaluating model performance through several error metrics, the paper aims to clarify how different forecasting paradigms behave under realistic ITS monitoring conditions. In doing so, it addresses the need for transparent and scalable evidence on the comparative value of time-series and machine-learning models for predictive pavement analytics using connected vehicles data.

2. Materials and Methods

NIRA Dynamics AB provided the dataset used in this study, derived from a fleet of connected vehicles equipped with original equipment manufacturer (OEM) sensors. The company currently operates a network of more than 2 million vehicles across Europe and North America, continuously collecting information on road pavement conditions. In particular, the dataset included an IRI-related roughness estimation supplied directly by NIRA Dynamics AB. While this indicator is not derived from the standard IRI calculation procedure based on profilometric measurements, it represents a reliable surrogate for pavement roughness. To support this the substantial body of literature reports significant correlation between connected-vehicle-based response measures and conventionally measured IRI. For instance, in a study of 2022, Mahlberg et al., [23] conducted a comparative study between IRI values obtained from inertial profilers and crowdsourced FCD data, employing a linear correlation model. The results of this study highlight a significant linear correlation between the two clusters ($R^2 = 0.79$ and p-value < 0.001), thereby substantiating their reliability and validity.

To avoid confusing the reader, the term "IRI" will be replaced in the manuscript with "crowdsourced_IRI," which refers specifically to the surrogate identifier provided by the provider.

Data acquisition, aggregation and anonymization were performed by NIRA Dynamics AB in compliance with the General Data Protection Regulation and current European Union regulations. The connected-vehicle data considered in this work were recorded daily from 21 September 2021 to 23 June 2024.

The proposed methodology was implemented in Python (v. 5.12) and developed as a unified benchmarking framework for daily road surface roughness forecasting at the road-segment level using FCD observations. The methodological objective was not limited to forecast generation alone, but rather to provide a transparent and reproducible basis for comparing different predictive paradigms under identical experimental conditions. Specifically, the framework was designed to assess if daily pavement roughness can be more effectively predicted through a classical seasonal time-series model or supervised machine-learning approaches driven by temporal, reliability-related and calendar-based predictors. The motorway network presented here has a total length of approximately 130 km, which is further subdivided into five sections. This network is hereinafter referred to as Route_ID. Table 1 presents the reference length of each section and the number of lanes. It is imperative to note that the number of lanes is a crucial factor in this configuration, as the crowdsourced_IRI data is presented in discrete units of measurement termed 'sub_segment_id', with a length of 25 m; in addition, the roads are geometrically subdivided into 'segment_id', where each of them is made of a certain number of sub_segments and their length depends on characteristics of the single road and of the element (i.e., junction, road between junctions, etc.). However, in the case of a road with multiple lanes, no

distinction is made between them and an overall crowdsourced_IRI value is provided for the entire section, which is 25 m long and of variable width.

Table 1. Road-segment description.

Segment_ID	Number of Sub-Segments of 25 m Length	Number of Lanes
1	25	3
2	25	3
3	20	2
4	30	4
5	30	4

The workflow comprised four main stages: (i) loading and harmonizing multi-file monitoring data, (ii) constructing regular daily time series for each road segment, (iii) transforming each daily series into a supervised forecasting dataset through engineered predictors, and (iv) comparing competing models through both unweighted and reliability-aware weighted evaluation metrics. The workflow is visually represented in Figure 1 and this structure was selected to ensure methodological consistency across heterogeneous road segments and to support computationally scalable application to large connected-vehicle archives.

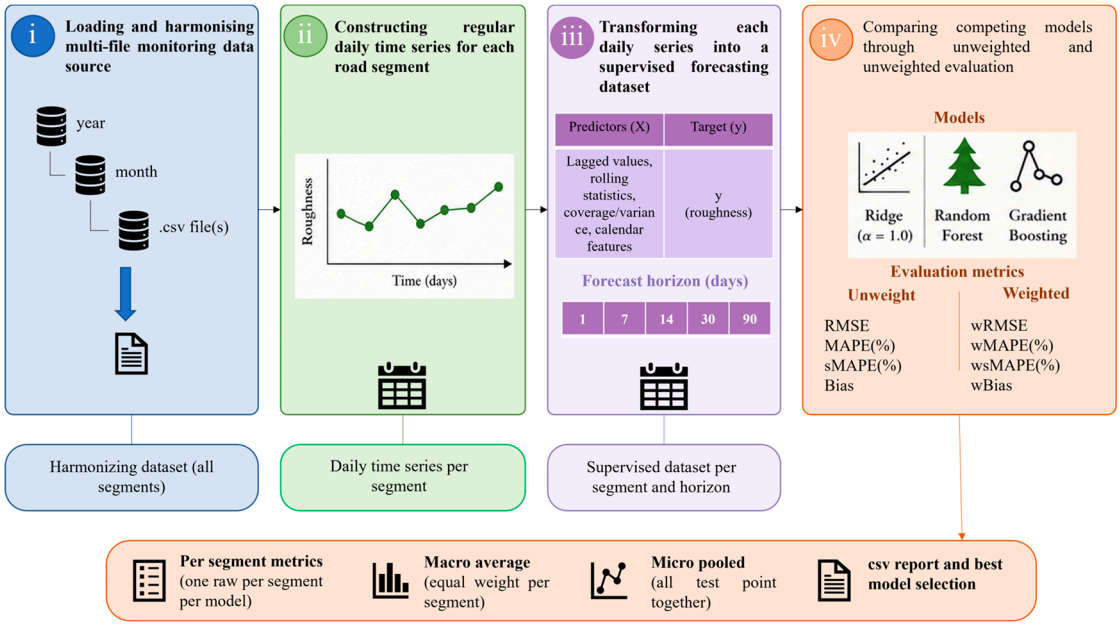

Figure 1. Methodological workflow representation.

2.1. Forecasting Formulation

The forecasting task was formulated as a direct multi-horizon daily prediction problem and was conducted independently for each road segment. Let $y_{i,t}$ denote the daily roughness value associated with road segment (i) at forecast origin (t). For each forecasting horizon h = (1, 7, 14, 30, 90) days, the objective was to estimate the future roughness value $\hat{y}_{i,t+h}$ using only information available at or before the forecast origin. A separate supervised forecasting dataset was therefore constructed for each road segment and forecasting horizon.

For each segment, the daily series was ordered chronologically and divided into training and test subsets using an 80/20 temporal split, which is a commonly adopted splitting ratio in the forecasting literature [24]. More specifically, the first 80% of the daily series defined the calibration period, whereas the remaining 20% defined the out-of-sample validation period. To prevent temporal leakage in the multi-horizon setting, training cases

were retained only when their corresponding target date ($t + h$) fell within the training period. Conversely, test cases were defined as those for which the target date fell within the holdout period. This procedure reproduced a realistic forecasting setting in which future pavement conditions were predicted exclusively from previously available information.

Before constructing the supervised datasets, only road segments containing at least 180 daily observations were considered (no segments were discarded). After feature generation, an additional consistency check was applied separately for each forecasting horizon and only segment–horizon combinations containing at least 50 supervised training samples and 10 supervised test samples were retained (no segments were discarded). These thresholds were introduced to reduce the risk of unstable model calibration and unreliable performance evaluation for short effective time series.

For the machine-learning models, the target variable of each supervised dataset was the roughness value $y_{i,t+h}$, while the explanatory variables described the condition and reliability of the available observations at the forecast origin. Since a direct forecasting strategy was adopted, each horizon was treated as an independent prediction problem rather than being obtained recursively from shorter-term predictions.

The predictor space was designed to represent multiple dimensions of roughness evolution. The first family consisted of autoregressive predictors derived from roughness values lagged by 1, 2, 3, 7, 14, 21, 28, 56, and 90 days. These variables were introduced to represent short-, medium-, and longer-term temporal dependence under the assumption that the recent history of the pavement-condition indicator contains information relevant to its future evolution.

The second predictor family comprised rolling-window statistics calculated exclusively from historical roughness observations. Specifically, the 7-day and 28-day rolling means and standard deviations were included. All rolling statistics were temporally shifted before their calculation so that the value being predicted could not contribute to its own predictors. These descriptors summarized the recent local level and variability of the roughness signal, thereby extending the information provided by individual lagged observations.

A third predictor family represented data availability and observation support through the number of connected-vehicle observations contributing to the daily estimate. When available, the current coverage value at the forecast origin, its values lagged by 1, 7, 14, and 28 days, and its 7-day and 28-day rolling means were included. These variables allowed the models to account for temporal changes in the empirical support underlying the daily roughness estimate.

A fourth predictor family described estimation uncertainty. The current roughness-state variance, its values lagged by 1, 7, 14, and 28 days, and its 7-day and 28-day rolling means were included when available. Additional state-estimation variables, including the estimated roughness level and its associated variance, were also represented through their current values and lagged values at 1, 7, 14, and 28 days. These predictors were introduced because connected-vehicle-derived roughness observations may have similar nominal values while differing substantially in their statistical reliability.

Coverage and variance were therefore used in two complementary ways. First, they were included among the predictors so that the forecasting models could learn whether the reliability and availability of the observations were associated with future roughness values. Second, they were used to derive the uncertainty-aware sample weights employed during model calibration and weighted performance evaluation.

Finally, calendar predictors were included to represent possible periodic patterns in connected-vehicle data availability and roughness estimation. These consisted of the day of the week, month of the year, and a binary weekend indicator. Although these variables

do not directly describe pavement deterioration, they may capture systematic temporal variations in traffic coverage and measurement conditions.

The resulting predictor families and the direct multi-horizon forecasting formulation are summarized in Figure 2.

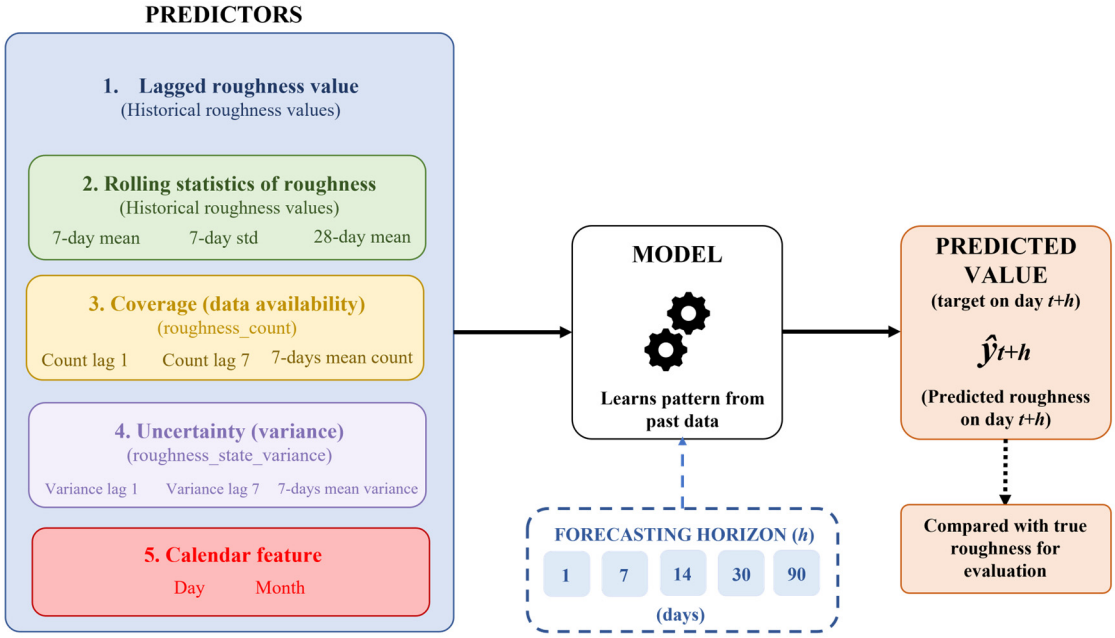

Figure 2. Visual representation of predictors and predicted variable.

Once the predictors were defined, three forecasting approaches were benchmarked:

- Ridge Regression;
- Random Forest (RF);
- Gradient Boosting (GB).

These models were selected to represent distinct modeling philosophies and increasing levels of flexibility.

Ridge regression was selected as a regularized linear model capable of handling multicollinearity among lagged and rolling predictors while preserving stability and interpretability. Random Forest and Gradient Boosting were instead adopted as nonlinear methods capable of modeling more complex interactions and non-additive relationships among predictors. Model hyperparameters were selected separately for each road segment and forecasting horizon using grid-search cross-validation. To preserve the chronological structure of the observations, an expanding-window time-series cross-validation procedure with five sequential folds was applied exclusively to the training dataset. The final 20% chronological holdout period was excluded from hyperparameter selection and retained only for out-of-sample performance evaluation.

The candidate configurations were evaluated using RMSE as the cross-validation criterion. For Ridge regression, the regularization parameter α was evaluated over the values 0.01, 0.1, 1, 10, and 100. For Random Forest, the search considered 100, 300, or 500 trees; maximum depths of 3, 5, 10, or no predefined limit; and minimum leaf sizes of 1, 5, or 10 observations. For Gradient Boosting, the investigated configurations included 100 or 300 estimators; learning rates of 0.03, 0.05, or 0.1; maximum depths of 2, 3, or 5; and minimum leaf sizes of 1, 5, or 10 observations. After identifying the configuration with the lowest mean cross-validation RMSE, the corresponding model was refitted using the complete training dataset and evaluated on the untouched holdout period. A fixed random seed was adopted for the ensemble models solely to ensure reproducibility.

A key methodological component of the framework was the adoption of a sample-weighting strategy for model fitting and performance assessment. The implemented methodology allowed three alternative weighting formulations based on daily observation count, inverse variance or a combination of both. In the current configuration, the selected scheme was the combined weighting strategy, defined as Equation (1):

$$w_{i,t} = \frac{\sqrt{c_{i,t}}}{v_{i,t} + \varepsilon}, \quad (1)$$

where $c_{i,t}$ denotes the daily number of contributing observations, $v_{i,t}$ denotes the daily state-estimation variance, and $\varepsilon = 10^{-6}$ is a small numerical stabilization term. The resulting weights were then normalized to the unit mean before use.

The rationale for this weighting strategy lies in the nature of connected-vehicle monitoring data. Unlike conventional profilometric surveys, probe-based estimates are generated under varying observation support and heterogeneous confidence conditions.

The introduction of weighted performance metrics was motivated by the intrinsic heterogeneity of connected-vehicle observations. Unlike traditional pavement surveys, where each observation is typically acquired under controlled and standardized conditions, connected-vehicle-based roughness estimates are derived from varying numbers of vehicle passages and may therefore exhibit different levels of statistical reliability. In particular, days characterized by a large number of contributing observations are expected to provide a more robust estimate of the underlying pavement condition, whereas days with limited vehicle coverage are inherently more uncertain. Similarly, the state variance associated with the roughness estimate can be interpreted as an indicator of estimation uncertainty, with larger variance values reflecting lower confidence in the inferred road condition. Consequently, treating all observations as equally reliable may lead both model training and performance assessment to be disproportionately influenced by noisy or weakly supported measurements.

To account for this aspect, an uncertainty-aware weighting strategy was introduced. The adopted weighting scheme assigns greater importance to observations supported by a larger number of connected-vehicle measurements and lower estimated uncertainty, while reducing the influence of observations characterized by sparse data availability or high variance. The objective is not to alter the underlying prediction task, but rather to evaluate forecasting performance in a manner that better reflects the confidence associated with each roughness estimate. This approach is consistent with reliability-based data fusion and uncertainty-aware modeling principles, where observations associated with higher confidence are given greater influence during both model calibration and evaluation.

2.2. Performance Evaluation

Model performance was assessed through both segment-wise and pooled analyses. For each segment and each model, standard error metrics were computed, namely: Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), symmetric Mean Absolute Percentage Error (sMAPE) and Bias. Their weighted counterparts, namely wRMSE, wMAPE, wsMAPE and wBias, were also calculated using the reliability-aware sample weights, which are described above. These metrics were selected because RMSE is more sensitive to large errors, MAPE expresses prediction error on a relative scale, sMAPE provides a more symmetric percentage-based error measure while Bias identifies systematic overestimation (if Bias > 0) or underestimation (if Bias < 0) tendencies. Unlike RMSE, Bias is not a pure magnitude-based error metric because positive and negative residuals may compensate each other; it was therefore used as a complementary diagnostic indicator

of systematic model tendency. For the model evaluation, the unweighted metrics were computed as Equations (2)–(5).

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}, \tag{2}$$

$$\text{MAPE} = \frac{100}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{d_t}\right|, \tag{3}$$

$$\text{sMAPE} = \frac{100}{N}\sum_{t=1}^{N}\frac{2|\hat{y}_t - y_t|}{|y_t| + |\hat{y}_t| + \varepsilon} \tag{4}$$

$$\text{Bias} = \frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t - y_t) \tag{5}$$

where $d_t = y_t$ when $|y_t| \geq 10^{-9}$ and $d_t = 10^{-9}$ otherwise, in order to avoid numerical instability when the observed approaches zero; y_t, is the observed roughness value at day t, while $\hat{y}_t$ is the forecasted roughness value at day t.

The weighted metrics were introduced to account for the heterogeneous reliability of connected-vehicle-derived roughness observations. In this context, not all daily roughness estimates are equally informative, since some values are supported by a larger number of vehicle passages and lower estimation uncertainty, whereas others may be based on sparse observations or higher variance. The uncertainty-aware weighting strategy therefore assigns greater influence to observations characterized by higher reliability, while reducing the influence of noisier or less supported measurements.

The weighted metrics were computed as Equations (6)–(9).

$$\text{wRMSE} = \sqrt{\frac{\sum_{t=1}^{N} w_t (y_t - \hat{y}_t)^2}{\sum_{t=1}^{N} w_t}}, \tag{6}$$

$$\text{wMAPE} = 100\frac{\sum_{t=1}^{N} w_t \left|\frac{y_t - \hat{y}_t}{d_t}\right|}{\sum_{t=1}^{N} w_t}, \tag{7}$$

$$\text{wsMAPE} = 100\frac{\sum_{t=1}^{N} w_t \frac{2|\hat{y}_t - y_t|}{|y_t| + |\hat{y}_t| + \varepsilon}}{\sum_{t=1}^{N} w_t}, \tag{8}$$

$$\text{wBias} = \frac{\sum_{t=1}^{N} w_t (\hat{y}_t - y_t)}{\sum_{t=1}^{N} w_t}, \tag{9}$$

where (w_t) is the reliability-aware weight associated with observation (t). The weighted metrics were not intended to replace the unweighted evaluation, but to provide a complementary uncertainty-aware perspective on forecasting performance. In particular, wRMSE, wMAPE, and wsMAPE quantify prediction errors after assigning greater importance to more reliable observations, while wBias identifies whether systematic overestimation or underestimation persists under reliability-weighted evaluation.

Two complementary aggregation strategies were then adopted, namely MacroMetric (macro) and MicroMetric (micro), to summarize model performance across multiple road segments.

First, macro-level evaluation was obtained by computing the metrics separately for each segment (i) and then averaging them across all segments (S), thus assigning equal importance to every road segment regardless of its number of observations, as shown

in Equation (8). So, the usage of macro fits well when segments are characterized by heterogenous features and the objective is to avoid the dominance of large segments.

$$\text{MacroMetric} = \frac{1}{S}\sum_{i=1}^{S} Metric_i, \tag{10}$$

Second, micro-level evaluation was obtained by pooling all test observations across all segments (i) before computing the metrics, thereby assigning equal importance to each individual test sample (Equation (10)). The usage of micro is instead preferable for use when the objective is to assess the overall prediction accuracy and in the case of segments with heterogenous size.

$$\text{MicroMetric} = \text{Metric}(\bigcup_{i=1}^{S}\{y_{i,t}\}, \bigcup_{i=1}^{S}\{\hat{y}_{i,t}\}) \tag{11}$$

This dual perspective has been deliberately chosen to discriminate between segment-wise robustness and overall network-level predictive accuracy.

Overall, the described methodology has been designed to be transparent, scalable, and a reproducible benchmarking framework for daily road surface roughness forecasting based solely on crowdsourced_IRI data. Its main purpose was to determine whether future roughness trajectories are better captured by classical univariate time-series modeling or by supervised machine-learning approaches that integrate temporal dependence, recent variability, periodic effects, and reliability-related information. In practical terms, the methodology supports the selection of forecasting models for predictive pavement monitoring and provides a structured basis for integrating roughness forecasting into maintenance planning and decision-support processes.

3. Results

For the sake of clarity, the discussion of model performance primarily focuses on RMSE and wRMSE. RMSE was considered the main standard accuracy metric because it is expressed in the same unit as the roughness indicator and penalizes larger forecasting errors. The corresponding weighted metric, wRMSE, was used as a complementary reliability-aware indicator, since it assigns greater influence to observations characterized by higher connected-vehicle coverage and lower estimation uncertainty. The remaining metrics, including MAPE, sMAPE, and Bias, (and their weighted counterparts) were used as supporting indicators to verify the consistency of the observed model rankings.

However, Tables 1 and 2 provide an overview of all the metrics. It should be noted that the values for the data not explicitly mentioned follow the same trend as those analyzed. In particular, macro-level and micro-level results will be presented separately (Sections 3.1 and 3.2 respectively) because, as previously mentioned, they are two complementary aggregation strategies.

3.1. Macro Results

The macro-level evaluation showed a marked dependence of forecasting performance on the prediction horizon. The values reported in Table 2 represent the mean of the route-specific macro metrics across the five investigated routes, while the corresponding standard deviations reported in Table 3 quantify the between-route variability in model performance.

Ridge regression achieved the lowest mean macro RMSE at the 1-, 7-, 14-, and 30-day forecasting horizons. More specifically, its mean RMSE increased from (0.0066 ± 0.0009) at the 1-day horizon to (0.0275 ± 0.0111), (0.0386 ± 0.0089), and (0.0602 ± 0.0175) at the 7-, 14-, and 30-day horizons, respectively. The corresponding wRMSE values were nearly identical, amounting to (0.0066 ± 0.0009), (0.0275 ± 0.0112), (0.0385 ± 0.0091), and (0.0602 ± 0.0177).

This close agreement indicates that the uncertainty-aware weighting procedure did not substantially alter the model ranking. The superior short- and medium-term performance of Ridge was therefore maintained after assigning greater importance to observations characterized by higher connected-vehicle coverage and lower estimation uncertainty.

Prediction error increased substantially with the forecasting horizon. Considering the best-performing model at each horizon, the mean macro RMSE was approximately 4.2 times higher at 7 days, 5.9 times higher at 14 days, 9.1 times higher at 30 days, and 11.6 times higher at 90 days than at the 1-day horizon. This pattern supports the interpretation that one-day-ahead forecasting is strongly influenced by short-term temporal persistence, whereas longer horizons provide a more demanding assessment of the models' predictive capability.

The between-route standard deviations were also generally higher at medium and long forecasting horizons. For Ridge, the RMSE standard deviation increased from 0.0009 at 1 day to 0.0111 at 7 days, 0.0089 at 14 days, 0.0175 at 30 days, and 0.0182 at 90 days. Although this increase was not strictly monotonic, the larger variability observed at the longer horizons indicates that the performance of the linear model became less homogeneous across the investigated routes.

At the 30-day horizon, Ridge retained the lowest mean RMSE, although the differences among the three models became limited. Ridge achieved an RMSE of (0.0602 ± 0.0175), compared with (0.0638 ± 0.0108) for Gradient Boosting and (0.0642 ± 0.0125) for Random Forest. The corresponding wRMSE values were (0.0602 ± 0.0177), (0.0641 ± 0.0108), and (0.0644 ± 0.0125), respectively. Thus, while Ridge provided the lowest average error, the ensemble models showed lower between-route variability. This suggests that Ridge remained preferable on average, but its performance was less consistent across routes at this forecasting horizon.

At the 90-day horizon, the model ranking changed. Random Forest achieved the lowest mean macro RMSE, equal to (0.0764 ± 0.0050), and was closely followed by Gradient Boosting with (0.0771 ± 0.0058). Ridge produced a higher mean error of (0.0867 ± 0.0182) and showed substantially greater variability across routes. The corresponding wRMSE values followed the same pattern, with (0.0769 ± 0.0053) for Random Forest, (0.0778 ± 0.0060) for Gradient Boosting, and (0.0874 ± 0.0188) for Ridge. Since the difference between Random Forest and Gradient Boosting was small relative to their standard deviations, the two ensemble models should be regarded as broadly comparable at the 90-day horizon rather than as showing a clear superiority of one over the other.

The remaining error metrics were consistent with the RMSE-based interpretation. Ridge showed the lowest relative errors at the short and medium horizons, whereas Random Forest achieved the lowest MAPE and sMAPE values at 90 days. Bias values remained small relative to the corresponding RMSE values, although the ensemble models generally exhibited a slight tendency to underestimate the roughness indicator.

Overall, the macro-level results indicate that Ridge regression provides the lowest average segment-level errors at short and medium forecasting horizons, suggesting that much of the short-term evolution of the connected-vehicle-derived roughness indicator can be represented through regularized linear temporal relationships. However, its advantage decreased as the forecasting horizon increased, while the nonlinear ensemble models became more competitive and showed greater consistency across routes. Longer-term forecasting may therefore be more affected by route-specific dynamics and by explanatory factors not explicitly represented in the current predictor set, such as traffic loading, weather conditions, and maintenance history. Consequently, short-horizon forecasts appear more suitable for continuous monitoring and anomaly detection, whereas the use of longer-horizon forecasts for maintenance-oriented applications would require further validation and the integration of additional explanatory variables.

Table 2. Mean route-level macro performance metrics across the five investigated routes.

Horizon Days	Model	RMSE	MAPE%	sMAPE%	Bias	wRMSE	wMAPE%	wsMAPE%	wBias
1	Ridge	0.0066	0.2853	0.2851	0.0000	0.0066	0.2853	0.2851	0.0001
1	BG	0.0226	0.8700	0.8768	−0.0065	0.0224	0.8638	0.8702	−0.0062
1	RF	0.0253	0.9881	0.9955	−0.0071	0.0251	0.9811	0.9879	−0.0067
7	Ridge	0.0275	1.1432	1.1360	0.0019	0.0275	1.1441	1.1365	0.0023
7	BG	0.0413	1.7713	1.7717	−0.0044	0.0412	1.7678	1.7676	−0.0040
7	RF	0.0428	1.8259	1.8241	−0.0032	0.0426	1.8198	1.8174	−0.0028
14	Ridge	0.0386	1.6537	1.6395	0.0026	0.0385	1.6552	1.6404	0.0031
14	BG	0.0516	2.2539	2.2480	−0.0035	0.0515	2.2518	2.2452	−0.0030
14	RF	0.0519	2.2954	2.2848	−0.0010	0.0518	2.2902	2.2790	−0.0006
30	Ridge	0.0602	2.6423	2.5949	0.0044	0.0602	2.6481	2.6003	0.0048
30	BG	0.0638	2.8944	2.8767	−0.0070	0.0641	2.9144	2.8956	−0.0067
30	RF	0.0642	2.9213	2.9109	−0.0067	0.0644	2.9399	2.9288	−0.0064
90	Ridge	0.0867	3.8898	3.8236	−0.0027	0.0874	3.9215	3.8501	−0.0018
90	BG	0.0771	3.6217	3.5857	−0.0069	0.0778	3.6553	3.6186	−0.0071
90	RF	0.0764	3.5918	3.5569	−0.0051	0.0769	3.6155	3.5814	−0.0054

Table 3. Between-route sample standard deviation of the macro performance metrics across the five investigated routes.

Horizon Days	Model	RMSE	MAPE%	sMAPE%	Bias	wRMSE	wMAPE%	wsMAPE%	wBias
1	Ridge	0.0009	0.0397	0.0398	0.0010	0.0009	0.0388	0.0389	0.0011
1	BG	0.0077	0.1823	0.1851	0.0029	0.0076	0.1760	0.1784	0.0025
1	RF	0.0079	0.1793	0.1827	0.0038	0.0077	0.1732	0.1759	0.0033
7	Ridge	0.0111	0.2387	0.2279	0.0075	0.0112	0.2455	0.2340	0.0078
7	BG	0.0123	0.3786	0.3782	0.0036	0.0122	0.3715	0.3710	0.0034
7	RF	0.0117	0.2807	0.2759	0.0027	0.0117	0.2754	0.2697	0.0028
14	Ridge	0.0089	0.1768	0.1673	0.0111	0.0091	0.1832	0.1720	0.0114
14	BG	0.0123	0.3786	0.3782	0.0036	0.0122	0.3715	0.3710	0.0034
14	RF	0.0112	0.4264	0.4218	0.0045	0.0111	0.4176	0.4131	0.0042
30	Ridge	0.0175	0.2455	0.2320	0.0192	0.0177	0.2538	0.2370	0.0196
30	BG	0.0108	0.4938	0.4994	0.0037	0.0108	0.4897	0.4938	0.0033
30	RF	0.0125	0.5147	0.5197	0.0032	0.0125	0.5142	0.5179	0.0029
90	Ridge	0.0182	0.3838	0.3814	0.0235	0.0188	0.4108	0.3976	0.0238
90	BG	0.0058	0.7443	0.7669	0.0117	0.0060	0.7348	0.7576	0.0115
90	RF	0.0050	0.9057	0.8973	0.0149	0.0053	0.8889	0.8815	0.0151

3.2. Micro Results

The micro-level evaluation also showed a marked dependence of model performance on the forecasting horizon. For each route, the micro metrics were calculated by pooling all test observations across the corresponding road segments before computing the performance indicators. The values reported in Table 4 therefore represent the mean of the five route-specific micro metrics, while the standard deviations reported in Table 5 quantify the variability in pooled forecasting performance across routes.

Ridge regression achieved the lowest mean micro RMSE at the 1-, 7-, and 14-day forecasting horizons. Its mean RMSE increased from (0.0073 ± 0.0015) at the 1-day horizon to (0.0285 ± 0.0109) and (0.0492 ± 0.0162) at the 7- and 14-day horizons, respectively. The corresponding wRMSE values were (0.0060 ± 0.0015), (0.0285 ± 0.0105), and (0.0492 ± 0.0167). These results confirm that Ridge regression was particularly effective at capturing the short-term temporal structure of the connected-vehicle-derived roughness signal. Its advantage at these horizons was also preserved after accounting for differences in observation reliability.

Table 4. Mean route-level micro performance metrics across the five investigated routes. For each route, the micro metrics were calculated by pooling all test observations across road segments before averaging the five route-specific values.

Horizon Days	Model	RMSE	MAPE%	sMAPE%	Bias	wRMSE	wMAPE%	wsMAPE%	wBias
1	Ridge	0.0073	0.2377	0.2376	0.0000	0.0060	0.2378	0.2376	0.0001
1	BG	0.0315	0.8700	0.8768	−0.0065	0.0311	0.8638	0.8702	−0.0062
1	RF	0.0339	0.9881	0.9955	−0.0071	0.0335	0.9811	0.9879	−0.0067
7	Ridge	0.0285	1.0267	1.0233	0.0005	0.0285	1.0272	1.0236	0.0009
7	BG	0.0533	1.7713	1.7717	−0.0044	0.0531	1.7678	1.7676	−0.0040
7	RF	0.0541	1.8259	1.8241	−0.0032	0.0539	1.8198	1.8174	−0.0028
14	Ridge	0.0492	1.6537	1.6395	0.0026	0.0492	1.6552	1.6404	0.0031
14	BG	0.0653	2.2539	2.2480	−0.0035	0.0651	2.2518	2.2452	−0.0030
14	RF	0.0643	2.2954	2.2848	−0.0010	0.0641	2.2902	2.2790	−0.0006
30	Ridge	0.0912	2.6423	2.5949	0.0044	0.0914	2.6481	2.6003	0.0048
30	BG	0.0853	3.0942	3.0768	−0.0090	0.0856	3.1127	3.0952	−0.0089
30	RF	0.0871	3.2197	3.2136	−0.0095	0.0873	3.2384	3.2319	−0.0093
90	Ridge	0.1325	3.7286	3.6542	0.0002	0.1344	3.7597	3.6810	0.0008
90	BG	0.1060	3.6217	3.5857	−0.0069	0.1071	3.6553	3.6186	−0.0071
90	RF	0.1033	3.5918	3.5569	−0.0051	0.1041	3.6155	3.5814	−0.0054

Table 5. Between-route sample standard deviation of the route-level micro performance metrics across the five investigated routes.

Horizon Days	Model	RMSE	MAPE%	sMAPE%	Bias	wRMSE	wMAPE%	wsMAPE%	wBias
1	Ridge	0.0015	0.0397	0.0398	0.0010	0.0015	0.0388	0.0389	0.0011
1	BG	0.0118	0.1823	0.1851	0.0029	0.0115	0.1760	0.1784	0.0025
1	RF	0.0109	0.1793	0.1827	0.0038	0.0105	0.1732	0.1759	0.0033
7	Ridge	0.0109	0.1793	0.1827	0.0038	0.0105	0.1732	0.1759	0.0033
7	BG	0.0177	0.1857	0.1880	0.0033	0.0178	0.1806	0.1820	0.0029
7	RF	0.0227	0.2807	0.2759	0.0027	0.0228	0.2754	0.2697	0.0028
14	Ridge	0.0162	0.1768	0.1673	0.0111	0.0167	0.1832	0.1720	0.0114
14	BG	0.0232	0.3786	0.3782	0.0036	0.0233	0.3715	0.3710	0.0034
14	RF	0.0184	0.4264	0.4218	0.0045	0.0186	0.4176	0.4131	0.0042
30	Ridge	0.0533	0.2455	0.2320	0.0192	0.0539	0.2538	0.2370	0.0196
30	BG	0.0189	0.9185	0.9226	0.0054	0.0192	0.9126	0.9171	0.0054
30	RF	0.0104	0.5055	0.5152	0.0045	0.0107	0.5021	0.5125	0.0050
90	Ridge	0.0740	0.3595	0.2791	0.0193	0.0770	0.3881	0.2988	0.0200
90	BG	0.0242	0.7443	0.7669	0.0117	0.0252	0.7348	0.7576	0.0115
90	RF	0.0291	0.9057	0.8973	0.0149	0.0308	0.8889	0.8815	0.0151

Forecasting accuracy progressively decreased as the prediction horizon increased. Considering the best-performing model at each horizon, the mean micro RMSE was approximately 3.9 times higher at 7 days, 6.7 times higher at 14 days, 11.7 times higher at 30 days, and 14.2 times higher at 90 days than at the 1-day horizon. This substantial increase confirms that one-day-ahead forecasting is strongly supported by temporal persistence, whereas medium- and long-term prediction constitutes a more challenging assessment of the models' actual forecasting capability.

At the 30-day horizon, the model ranking changed. Gradient Boosting achieved the lowest mean micro RMSE, equal to (0.0853 ± 0.0189), followed by Random Forest with (0.0871 ± 0.0104) and Ridge with (0.0912 ± 0.0533). The corresponding wRMSE values showed the same ranking, being (0.0856 ± 0.0192), (0.0873 ± 0.0107), and (0.0914 ± 0.0539), respectively. Although Gradient Boosting provided the lowest mean error, the difference from Random Forest was limited. Moreover, Random Forest showed considerably lower between-route variability, suggesting more consistent pooled performance across the five

routes. By contrast, the large standard deviation associated with Ridge indicates that its performance at the 30-day horizon was highly dependent on the investigated route.

At the 90-day horizon, Random Forest achieved the lowest mean micro RMSE, equal to (0.1033 ± 0.0291), closely followed by Gradient Boosting with (0.1060 ± 0.0242). Ridge produced a substantially higher mean error of (0.1325 ± 0.0740). The reliability-weighted results confirmed the same pattern, with wRMSE values of (0.1041 ± 0.0308) for Random Forest, (0.1071 ± 0.0252) for Gradient Boosting, and (0.1344 ± 0.0770) for Ridge. The small difference between the two ensemble models relative to their between-route variability indicates that their performance should be considered broadly comparable at this horizon. Random Forest achieved the lowest average error, whereas Gradient Boosting showed slightly greater consistency across routes. Ridge was both less accurate and substantially more variable, indicating limited robustness for long-horizon forecasting.

The comparison between the macro- and micro-level evaluations provides additional insight into model behavior. The macro analysis identified Ridge as the best-performing model up to the 30-day horizon, whereas the micro analysis indicated that Gradient Boosting achieved the lowest average pooled error at 30 days. This difference reflects the distinct aggregation principles of the two evaluation strategies. Macro metrics assign equal importance to each road segment, while micro metrics give greater influence, within each route, to segments containing a larger number of test observations. Consequently, Ridge appears to provide stronger average segment-level robustness, whereas the ensemble methods better capture the overall distribution of observations at longer forecasting horizons.

The similarity between RMSE and wRMSE was generally high and the reliability-aware weighting procedure did not materially change the model rankings. This indicates that the main findings were not artificially driven by the weighting scheme. Instead, the weighted metrics provided a complementary assessment showing that the same performance patterns remained evident when greater importance was assigned to observations characterized by higher connected-vehicle coverage and lower estimation uncertainty.

Overall, the micro-level results indicate a progressive transition from the clear short-term superiority of Ridge regression toward the greater competitiveness of nonlinear ensemble models at medium and long horizons. Ridge appears particularly suitable for short-term monitoring applications, where temporal persistence provides substantial predictive information. Conversely, Gradient Boosting and Random Forest provided lower average errors and greater cross-route stability at the 30- and 90-day horizons. These findings suggest that longer-term roughness forecasting may involve more complex and route-dependent relationships that are not fully represented by a regularized linear model. Nevertheless, the direct use of these forecasts for maintenance planning would require further validation and the integration of additional explanatory variables related to traffic loading, weather conditions, and maintenance activities.

3.3. Comparison Between Macro and Micro Evaluation

The combined analysis of macro- and micro-level metrics provides complementary information regarding model performance. The macro metrics characterize average segment-level robustness by assigning equal importance to each road segment, whereas the micro metrics characterize pooled predictive accuracy within each route by giving greater influence to segments containing a larger number of test observations. Consequently, agreement between the two approaches indicates that a model performs consistently across both individual segments and the overall distribution of observations, while differences in ranking reveal sensitivity to segment size and heterogeneity.

At the 1-, 7-, and 14-day horizons, the macro- and micro-level evaluations provided the same model ranking, with Ridge regression consistently achieving the lowest RMSE

and wRMSE. At the 1-day horizon, Ridge obtained a macro RMSE of (0.0066 ± 0.0009) and a micro RMSE of (0.0073 ± 0.0015). At 7 days, the corresponding values were (0.0275 ± 0.0111) and (0.0285 ± 0.0109), while at 14 days they increased to (0.0386 ± 0.0089) and (0.0492 ± 0.0162), respectively. This agreement indicates that the short-term superiority of Ridge was not restricted to either small or observation-rich segments, but represented a consistent pattern across the investigated datasets. It also supports the interpretation that short-horizon forecasting is largely governed by regular temporal dependence that can be effectively represented by a regularized linear model.

Differences between macro and micro evaluations became more evident as the forecasting horizon increased. At the 30-day horizon, the macro analysis identified Ridge as the best-performing model, with an RMSE of (0.0602 ± 0.0175), compared with (0.0638 ± 0.0108) for Gradient Boosting and (0.0642 ± 0.0125) for Random Forest. Conversely, the micro analysis identified Gradient Boosting as the model with the lowest pooled RMSE, equal to (0.0853 ± 0.0189), followed by Random Forest with (0.0871 ± 0.0104) and Ridge with (0.0912 ± 0.0533). The substantially larger micro-level standard deviation obtained by Ridge indicates that its pooled performance was highly variable across routes, despite retaining the lowest average segment-level error. Therefore, Ridge appears to provide stronger average performance across individual segments, whereas the ensemble models are more effective and more stable when the complete distribution of test observations is considered.

At the 90-day horizon, the two aggregation strategies converged again in identifying Random Forest as the model with the lowest average RMSE. Random Forest achieved a macro RMSE of (0.0764 ± 0.0050) and a micro RMSE of (0.1033 ± 0.0291). Gradient Boosting showed closely comparable results, with corresponding values of (0.0771 ± 0.0058) and (0.1060 ± 0.0242). Ridge produced higher errors in both evaluations, reaching (0.0867 ± 0.0182) at the macro level and (0.1325 ± 0.0740) at the micro level. The considerably higher micro-level variability of Ridge confirms that its long-horizon performance was strongly dependent on route and segment characteristics. Although Random Forest achieved the lowest mean error, the small difference from Gradient Boosting relative to their standard deviations indicates that the two ensemble models should be considered broadly comparable at this horizon.

The relative-error indicators generally supported the RMSE-based interpretation. MAPE and sMAPE increased with the forecasting horizon for all models, confirming that the loss of accuracy at longer horizons was evident in both absolute and relative terms. The similarity between MAPE and sMAPE suggests that the overall interpretation was not strongly affected by the asymmetry of the conventional MAPE formulation. An exception emerged in the 30-day micro evaluation: although Gradient Boosting achieved the lowest RMSE, Ridge retained the lowest MAPE and sMAPE. This suggests that Ridge produced lower average proportional errors but was affected by a limited number of larger absolute deviations, which were penalized more strongly by RMSE. This difference further supports the use of RMSE as the primary metric while retaining percentage-based indicators as complementary diagnostics.

Bias values were small relative to the corresponding RMSE values, indicating that systematic overestimation or underestimation was limited compared with the overall forecasting error. Ridge showed approximately unbiased predictions at the 1-day horizon and a slightly positive Bias at the 7-, 14-, and 30-day horizons, indicating a modest tendency to overestimate the roughness indicator. By contrast, Gradient Boosting and Random Forest generally showed negative Bias values, particularly at the medium and long horizons, indicating a slight tendency to underestimate roughness. However, the associated standard

deviations demonstrate that these tendencies were not fully homogeneous across routes and should therefore be interpreted cautiously.

The weighted metrics followed the same overall patterns as their unweighted counterparts. In particular, the comparison between RMSE and wRMSE did not materially change the model ranking at any forecasting horizon. Similarly, wMAPE, wsMAPE, and wBias remained close to the corresponding unweighted values. This consistency indicates that the uncertainty-aware weighting scheme did not artificially drive the conclusions, but rather confirmed that the observed performance patterns persisted when greater importance was assigned to observations characterized by higher connected-vehicle coverage and lower estimation uncertainty.

Overall, the comparison between macro- and micro-level evaluations (shown in Table 6) shows that Ridge regression is the most robust model for short-term forecasting, both at the individual-segment level and across pooled observations. At intermediate horizons, however, the difference between the two aggregation strategies reveals that model performance becomes more sensitive to segment size and route heterogeneity. Ridge retains strong average segment-level performance, whereas Gradient Boosting and Random Forest become more competitive when observation-rich segments exert greater influence. At the longest horizon, both evaluations favor nonlinear ensemble models, suggesting that longer-term roughness forecasting is less uniformly governed by linear temporal dependence and may require more flexible models together with additional explanatory variables.

Table 6. Best model by horizon.

Horizon	Best Macro Mode	Best Micro Mode
1 day	Ridge	Ridge
7 days	Ridge	Ridge
14 days	Ridge	Ridge
30 days	Ridge	GB
90 days	RF	RF

4. Discussion and Conclusions

In recent years, the literature has increasingly emphasized the limitations of traditional pavement survey methods. At the same time, the opportunities offered by connected-vehicle and floating-car-data systems for more continuous and extensive road condition monitoring are increasing. Within this broader evolution, forecasting is increasingly being investigated as a complementary monitoring tool, as it may help road agencies identify emerging changes in pavement-condition indicators and support planning activities when the forecasting horizon and level of validation are aligned with operational requirements [41].

The multi-horizon comparison showed that model performance depended strongly on the forecasting horizon. Ridge regression achieved the lowest average errors at the short forecasting horizons and remained the best-performing model in both the macro- and micro-level evaluations up to 14 days. At the 30-day horizon, Ridge retained the lowest macro-level error, whereas Gradient Boosting achieved the lowest micro-level error. At 90 days, Random Forest provided the lowest average error under both aggregation strategies, although its performance was broadly comparable with that of Gradient Boosting. These results suggest that regularized linear relationships are particularly effective for short-term forecasting, while nonlinear ensemble models become increasingly competitive as the prediction horizon increases. This interpretation concerns the predictive structure of the connected-vehicle-derived roughness signal and should not be taken as evidence that the

models reproduce the underlying physical mechanisms of pavement deterioration. The remaining performance indicators generally supported the same horizon-dependent pattern.

From a practical point of view, the results suggest that feature-based machine-learning models, and especially regularized linear approaches such as Ridge, can offer a solid basis for predictive pavement monitoring in ITSs. The proposed framework is relevant for real-world applications because it provides a transparent and scalable way to compare forecasting methods while also accounting for the variable reliability of connected-vehicle observations through weighted performance measures. The weighted evaluation produced results that were generally consistent with the unweighted metrics, suggesting that the observed model ranking was not merely driven by low-confidence observations. This finding supports the robustness of the proposed uncertainty-aware weighting strategy and indicates that the superior performance of the best-performing models was maintained even when greater importance was assigned to observations characterized by higher measurement reliability.

From an implementation perspective, the forecasting module could be connected to an agency's existing pavement management system or geographic information system. Newly available connected-vehicle observations could be aggregated at the road-segment level, processed automatically and used to update the forecast trajectories at predefined intervals. A map-based dashboard could highlight segments for which the observed roughness-related signal departs substantially from the forecast, allowing technical staff to focus their attention on a limited number of potentially critical locations rather than reviewing the complete network manually.

In operational terms, such an approach could support maintenance prioritization, improve the early identification of critical deterioration patterns, and contribute to a more proactive and data-driven management of road infrastructure. However, the practical use of the forecasts should be interpreted in light of the forecasting horizon considered in this study. A one-day-ahead prediction is unlikely to directly trigger maintenance interventions, which are typically planned over substantially longer timeframes. Rather than serving as a standalone maintenance decision-support tool, short-term forecasts can be viewed as an intermediate component within a continuous pavement monitoring framework, helping agencies detect emerging anomalies, track rapid condition changes, and assess the short-term evolution of roughness indicators derived from connected-vehicle data. For direct maintenance planning applications, longer forecasting horizon would likely be required, and future research should investigate whether the proposed framework maintains its predictive capability under these more operationally relevant conditions.

Within maintenance-planning workflows, short-term forecasts could therefore operate as a preliminary screening layer rather than as a direct treatment-selection mechanism. A segment flagged by the forecasting system could first be checked against its current condition classification, recent inspection results, traffic importance, previous maintenance activities, and already scheduled works. If the forecast deviation is persistent and supported by other condition indicators, the segment could then be transferred to a higher-priority inspection list. Conversely, an isolated deviation associated with low vehicle coverage or high estimation uncertainty could be retained for continued observation without immediately triggering field activity. In this way, the forecasts could help agencies allocate inspection resources more selectively while preserving the role of engineering assessment in final maintenance decisions.

The operational relevance of the forecasts also depends on the prediction horizon. Short-term forecasts at 1, 7, and 14 days are unlikely to directly determine maintenance interventions, but they may provide useful support for continuous monitoring and short-term condition verification. In particular, the predicted trajectory can serve as a model-

based reference against which newly observed roughness values are compared. Substantial deviations between the forecast and the observed trend may indicate an unexpected change in pavement condition, a possible anomaly in the connected-vehicle data, or the need for additional inspection.

This comparison may also be useful after a maintenance intervention. Once the intervention has been completed, the observed short-term roughness evolution could be compared with the forecast-based reference trajectory to assess whether the measured response is consistent with the anticipated improvement or stabilization of the roughness indicator. Similarly, the framework could support the early identification of rapid post-intervention deterioration or an unexpectedly limited improvement. Nevertheless, such an application should be interpreted as short-term performance monitoring rather than as a causal assessment of maintenance effectiveness, unless maintenance records and an appropriate counterfactual validation framework are explicitly incorporated.

More generally, 1-day forecasts may support anomaly screening and data-quality control, while 7- and 14-day forecasts may help confirm whether an observed change represents a persistent trend rather than an isolated fluctuation. These horizons could also support the prioritization of targeted inspections following maintenance activities, extreme weather events, sudden traffic changes or abrupt variations in the monitored roughness signal. Forecasts at 30 and 90 days are closer to condition-management timescales, but their direct use for intervention selection, budgeting or treatment scheduling would still require integration with field inspections, maintenance history, condition thresholds, traffic and environmental variables, economic constraints, and agency-specific decision rules. The proposed framework should therefore be regarded as evidence of forecasting feasibility and as a potential component of a continuous monitoring system, rather than as a fully validated maintenance decision-support tool.

At the same time, several limitations should be acknowledged.

First, the analysis was conducted on five routes belonging to a specific real-world case study, and the transferability of the results should therefore be evaluated on larger and more diverse road networks.

Second, the predicted variable was a connected-vehicle-derived roughness indicator rather than IRI measured through standard profilometric surveys. Although previous studies support its relationship with pavement roughness, the present analysis does not provide direct validation against conventional pavement-condition measurements or against observed physical deterioration mechanisms.

Third, although the revised framework considered forecasting horizons ranging from 1 to 90 days, even the longest horizon may remain shorter than the timescales typically required for rehabilitation programming, budgeting, and long-term asset management.

Fourth, the models were mainly based on historical roughness and coverages, while potentially relevant explanatory variables such as traffic loading, weather conditions, pavement structure, construction characteristics, and maintenance history were not included due to lack of data. Route-specific data availability and the procedures used to complete the daily time series may also influence model performance and should be examined through additional sensitivity analyses. In this regard, recent explainable-AI approaches demonstrate the potential of feature-attribution methods for pavement roughness prediction [41]. Such analyses represent a relevant future extension of the present framework, particularly because the current predictor set is dominated by correlated lagged and rolling temporal variables and does not include several physical deterioration factors.

Fifth, a formal characterization of the autocorrelation, stationarity, spectral structure, and effective complexity of the roughness time series was beyond the scope of the present study and should be investigated in future work.

Future research should therefore validate the forecasts against independent profilometric and maintenance datasets, investigate longer and operationally relevant prediction horizons, assess the sensitivity of the results to the weighting and data-completion procedures, and integrate additional explanatory variables. Interpretability analyses should also be used to identify the variables on which the models rely, while avoiding causal interpretations of purely predictive associations. Despite these limitations, the study demonstrates that multi-year connected-vehicle observations contain exploitable temporal information for forecasting the future evolution of a roughness-related signal. The results support the potential use of such forecasts as a component of continuous pavement monitoring systems, while their direct application to maintenance decision-making requires further external and operational validation.

Author Contributions: Conceptualization, R.C., M.P. and L.C.; methodology, R.C., V.V. and C.L.; software, R.C. and L.C.; validation, L.C., V.V. and C.L.; formal analysis, R.C., M.P. and L.C.; investigation, R.C., M.P. and L.C.; data curation, R.C. and L.C.; writing—original draft preparation, R.C. and M.P.; writing—review and editing, V.V., L.C. and C.L.; supervision, C.L. 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 datasets presented in this article are not readily available because the data are part of an ongoing study Requests to access the datasets should be directed to Riccardo Ceriani.

Acknowledgments: This study was carried out within the MOST–SustainableMobility National Research Center and received financial support from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)–MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4–D.D. 1033 17/06/2022, CN00000023). This manuscript reflects only the authors' views and opinions, and neither the European Union nor the European Commission can be considered responsible for them. Authors would like to thanks NIRA Dynamics for providing the data within the above-mentioned project.

Conflicts of Interest: Authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ITSs	Intelligent Transport Systems
FDC	Floating Car Data
IRI	International Roughness Index
GB	Gradient Boosting
RF	Random Forest
RMSE	Route Mean Square Error
wRMSE	Weighted Route Mean Square Error
MAPE	Mean Absolute Percentage Error
sMAPE	symmetric Mean Absolute Percentage Error
wMAPE	weighted Mean Absolute Percentage Error
wsMAPE	weighted symmetric Mean Absolute Percentage Error

References

1. Famewo, B.G.; Shokouhian, M. A Review of Pavement Performance Deterioration Modeling: Influencing Factors and Techniques. *Symmetry* **2025**, *17*, 1992. [CrossRef]
2. Hasibuan, G.C.R.; Anas, M.R.; Dalimunthe, N.I.P.; Fath, M.T.A.; Dewi, R.A.; Syafridon, G.G.A.; Jaya, I.; Syahrizal. A Global Perspective on Road Condition Assessment and Maintenance: Trends in Research and Technology Integration. *Transp. Res. Interdiscip. Perspect.* **2026**, *36*, 101853. [CrossRef]
3. Pérez, A.; Sánchez, C.N.; Velasco, J. Damage Importance Analysis for Pavement Condition Index Using Machine-Learning Sensitivity Analysis. *Infrastructures* **2024**, *9*, 157. [CrossRef]
4. Li, Y.; Liu, C.; Ding, L. Impact of pavement conditions on crash severity. *Accid. Anal. Prev.* **2013**, *59*, 399–406. [CrossRef] [PubMed]
5. Amani, M.J.; Golroo, A.; Hajikarimi, P. Surrogate measurement of traffic safety with insights from pavement condition, sensor fusion, and machine learning. *Sci. Rep.* **2025**, *16*, 1862. [CrossRef] [PubMed]
6. Chen, S.; Saeed, T.U.; Labi, S. Impact of road-surface condition on rural highway safety: A multivariate random parameters negative binomial approach. *Anal. Methods Accid. Res.* **2017**, *16*, 75–89. [CrossRef]
7. Allen, E.; Costello, S.B.; Henning, T.F.P. Quantifying road criticality through the impact of disruptions on pavement deterioration and agency maintenance costs. *Int. J. Disaster Risk Reduct.* **2025**, *126*, 105592. [CrossRef]
8. Tian, G.; Jia, Y.; Chen, Z.; Gao, Y.; Wang, S.; Wei, Z.; Chen, Y.; Zhang, T. Evaluation on Lateral Stability of Vehicle: Impacts of Pavement Rutting, Road Alignment, and Adverse Weather. *Appl. Sci.* **2023**, *13*, 3250. [CrossRef]
9. Wu, X.; Chen, Q.; Li, Y.; Dong, N.; Yu, H. Investigating the Deterioration of Pavement Skid Resistance Using an Accelerated Pavement Test. *Materials* **2023**, *16*, 422. [CrossRef] [PubMed]
10. Alhasan, A.; Nlenanya, I.; Smadi, O.; MacKenzie, C.A. Impact of Pavement Surface Condition on Roadway Departure Crash Risk in Iowa. *Infrastructures* **2018**, *3*, 14. [CrossRef]
11. Hashim, I.H.; Badawy, R.M.; Heneash, U. Impact of Pavement Defects on Traffic Operational Performance. *Sustainability* **2023**, *15*, 8293. [CrossRef]
12. Llopis-Castelló, D.; Arce-Blanco, C.; Valdecantos-Álvarez, J.C. Before–after study of pavement roughness impact on vehicle fuel consumption and emissions. *Transp. Res. Part D Transp. Environ.* **2026**, *151*, 105131. [CrossRef]
13. Ajayi, O.O.; Kurien, A.M.; Djouani, K.; Dieng, L. Analysis of Road Roughness and Driver Comfort in 'Long-Haul' Road Transportation Using Random Forest Approach. *Sensors* **2024**, *24*, 6115. [CrossRef] [PubMed]
14. Misaghi, S.; Tirado, C.; Nazarian, S.; Carrasco, C. Impact of pavement roughness and suspension systems on vehicle dynamic loads on flexible pavements. *Transp. Eng.* **2021**, *3*, 100045. [CrossRef]
15. Wu, D. Ride Comfort Prediction on Urban Road Using Discrete Pavement Roughness Index. *Appl. Sci.* **2024**, *14*, 3108. [CrossRef]
16. Zaabar, I. Effect of Pavement Condition on Vehicle Operating Costs Including Fuel Consumption, Vehicle Durability and Damage to Transported Goods. Ph.D. Thesis, Michigan State University, East Lansing, MI, USA, 2010.
17. Ahmed, S.; Vedagiri, P.; Krishna Rao, K.V. Prioritization of pavement maintenance sections using objective based Analytic Hierarchy Process. *Int. J. Pavement Res. Technol.* **2017**, *10*, 158–170. [CrossRef]
18. Farhan, J.; Fwa, T.F. Use of Analytic Hierarchy Process to Prioritize Network-Level Maintenance of Pavement Segments with Multiple Distresses. *Transp. Res. Rec. J. Transp. Res. Board* **2011**, *2225*, 11–20. [CrossRef]
19. Wang, L.B.; Park, J.; Hill, S.H. Use of Pavement Management System Data to Monitor Performance of Pavements under Warranty. *Transp. Res. Rec. J. Transp. Res. Board* **2005**, *1940*, 21–31. [CrossRef]
20. Fares, A.; Zayed, T. Industry- and Academic-Based Trends in Pavement Roughness Inspection Technologies over the Past Five Decades: A Critical Review. *Remote Sens.* **2023**, *15*, 2941. [CrossRef]
21. Byrne, M.; Albrecht, D.; Sanjayan, J. Identifying error and maintenance intervention of pavement roughness time series with minimum message length inference. *Int. J. Pavement Eng.* **2010**, *11*, 37–47. [CrossRef]
22. Abdelaziz, N.; Abd El-Hakim, R.T.; El-Badawy, S.M.; Afify, H.A. International Roughness Index prediction model for flexible pavements. *Int. J. Pavement Eng.* **2018**, *21*, 88–99. [CrossRef]
23. Múčka, P. International Roughness Index specifications around the world. *Road Mater. Pavement Des.* **2016**, *18*, 929–965. [CrossRef]
24. *ASTM E1926-08*; Standard Practice for Computing International Roughness Index of Roads from Longitudinal Profile Measurements. ASTM International: West Conshohocken, PA, USA, 2021.
25. Joseph, V.R. Optimal ratio for data splitting. *Stat. Anal. Data Min. ASA Data Sci. J.* **2022**, *15*, 531–538. [CrossRef]
26. Sayers, M.W.; Gillespie, T.D.; Queiroz, C.A.V. *The International Road Roughness Experiment: Establishing Correlation and a Calibration Standard for Measurements*; World Bank Technical Paper; World Bank: Washington, DC, USA, 1986.
27. Fiorentini, N.; Losa, M. Road Detection, Monitoring, and Maintenance Using Remotely Sensed Data. *Remote Sens.* **2025**, *17*, 917. [CrossRef]
28. Sharma, M.; Al-Hammadi, M.; Mork, H.; Klein-Paste, A. Condition Assessment of Low-Volume Roads in Norway: A Deep Learning Approach for Automatic Assessment at Network Level. *Int. J. Pavement Res. Technol.* **2025**.

29. Pierce, L.M.; McGovern, G.; Zimmerman, K.A. *Practical Guide for Quality Management of Pavement Condition Data Collection*; Federal Highway Administration, U.S. Department of Transportation: Washington, DC, USA, 2013.
30. Ranyal, E.; Sadhu, A.; Jain, K. Road Condition Monitoring Using Smart Sensing and Artificial Intelligence: A Review. *Sensors* **2022**, *22*, 3044. [CrossRef] [PubMed]
31. Husain, S.F.; Tutumluer, E.; Mechitov, K.A.; Qamhia, I.I.A.; Spencer, B.; Riley Edwards, J. Towards a wireless sensing infrastructure for smart mobility. *Transp. Geotech.* **2023**, *40*, 100985. [CrossRef]
32. Staniek, M. Road pavement condition diagnostics using smartphone-based data crowdsourcing in smart cities. *J. Traffic Transp. Eng. (Engl. Ed.)* **2021**, *8*, 554–567. [CrossRef]
33. Khanmohamadi, M.; Guerrieri, M. Advanced Sensor Technologies in CAVs for Traditional and Smart Road Condition Monitoring: A Review. *Sustainability* **2024**, *16*, 8336. [CrossRef]
34. Yiliguoqi, B.; Siriguleng, B.; Arong Guo, G.; Wang, J. Research on the evaluation and analysis of road surface roughness based on smartphone sensors and SVM. *Sci. Rep.* **2026**, *16*, 4409. [CrossRef] [PubMed]
35. Mazzi, C.; Carini, C.; Meocci, M.; Paliotto, A.; Marradi, A. Floating Car Data for Road Roughness: An Innovative Approach to Optimize Road Surface Monitoring and Maintenance. *Future Transp.* **2025**, *5*, 162. [CrossRef]
36. Ceriani, R.; Vignali, V.; Chiola, D.; Pazzini, M.; Pettinari, M.; Lantieri, C. Exploring the Effectiveness of Road Maintenance Interventions on IRI Value Using Crowdsourced Connected Vehicle Data. *Sensors* **2025**, *25*, 3091. [CrossRef] [PubMed]
37. Mahlberg, J.A.; Li, H.; Zachrisson, B.; Leslie, D.K.; Bullock, D.M. Pavement Quality Evaluation Using Connected Vehicle Data. *Sensors* **2022**, *22*, 9109. [CrossRef] [PubMed]
38. Al-Samahi, S.; Zeiada, W.; Al-Khateeb, G.G.; Hamad, K.; Alnaqbi, A. A Comparative Study of Pavement Roughness Prediction Models under Different Climatic Conditions. *Infrastructures* **2024**, *9*, 167. [CrossRef]
39. Guan, X.; Zhang, H.; Du, X.; Zhang, X.; Sun, M. An Improved Method for Optimizing the Timing of Preventive Maintenance of Pavement: Integrating LCA and LCCA. *Appl. Sci.* **2023**, *13*, 10629. [CrossRef]
40. Thompson, A.; Desai, J.; Bullock, D.M. Evaluation of Connected Vehicle Pavement Roughness Data for Statewide Needs Assessment. *Infrastructures* **2025**, *10*, 248. [CrossRef]
41. Erfani, A.; Shayesteh, N.; Adnan, T. Data-augmented explainable AI for pavement roughness prediction. *Autom. Constr.* **2025**, *176*, 106307. [CrossRef]