scientific reports

Check for updates

OPEN

Research on the evaluation and analysis of road surface roughness based on smartphone sensors and SVM

B. Yiliguoqi[1], B. Siriguleng[2]✉, Arong[3], Guoying Guo[1] & Jing Wang[1]

We present a cost-effective approach to evaluating road surface quality and roughness based on low-cost smartphones by leveraging accelerometer and gyroscope data sampled at 10 Hz. We extracted features from the vertical accelerations and rolling motions, in cloud to standard deviation and interquartile range, and trained a support vector machine (SVM) classifier to identify Good and Poor roughness levels based on IRI thresholds. We selected SVM specifically because it affords consistent performance with small datasets, reliably handles low-dimensional statistical features, and provides a stronger generalization than more complex machine-learning approaches. We provide experimental results from four vehicles on a thrice 50-m segment based dataset, demonstrating that the proposed method can discriminate (i.e., classify) roadway roughness and quality levels 80–100% of the time based on smartphone IMU data, which implies that even low-frequency smartphone IMU signals can provide useful roughness screening for planning and maintenance.

Keywords Maintenance, Support vector machine (SVM), Standard deviation, Simple diagnosis, Roughness

The International Roughness Index (IRI) is the most recognized universal standard for the assessment of pavement longitudinal roughness and is measured for the duration of is based on the total vertical displacement of a standardized vehicle model traveling at 80 km/h over the specified measured profile (IRI) is based on a vehicle model referred to as a quarter-car simulation, which is used to measure the total vertical displacement of a vehicle traveling 80 km/h over the measured profile, which is also based on is represented in units of meters per kilometer (m/km), which is also written in millimeters per meter (mm/m), with higher values relating to progressively rough pavement surfaces. Road agencies often generally designate threshold cut-off values, such as IRI < 4 mm/m for "Good," and IRI ≥ 4 mm/m for "Poor" for guiding maintenance priorities and repair recommendations. The IRI is a valuable reference point for the interpretation of smartphone based pavement roughness measurements, as it is strongly correlated with ride quality, vehicle vibrations and maintenance costs.

The After large-scale highway construction, the mileage of China's highways has basically met the needs of economic development. As of the end of 2022, the total length of highways in China was 5.36 million kilometers, with maintenance mileage accounting for 99.9% of the total length of highways. Among them, the rural road mileage is 4.54 million kilometers, accounting for 84.6% of the total mileage[1]. The main task of highway transportation development has transitioned from the era of road construction to the era of maintenance, and highway maintenance has become a "heavy and urgent" task in highway transportation development. The "14th Five Year Plan for the Development of Highway Maintenance and Management"[2] issued by the Ministry of Transport of China clearly proposes to "explore and promote new types of non-destructive testing equipment, develop and promote the application of economically efficient automated testing equipment. Rural roads will basically achieve full coverage of road condition automated testing.". It can be seen that the intelligence of road performance monitoring methods, including rural roads, will be a key focus of later development. Further strengthening the detection methods of road surfaces, promoting the scientific, modern, and systematic detection of road surfaces, will become one of the key contents of China's transportation infrastructure maintenance and management work. At present, China has basically achieved annual inspection of all highways and quality inspection of typical national and provincial trunk lines, and has established a relatively complete

[1]Department of Transport and Municipal Engineering, Inner Mongolia Technical University of Construction, Hohhot Inner Mongolia 010000, China. [2]College of Mechanics and Aeronautics, Inner Mongolia University of Technology, Hohhot Inner Mongolia 010000, China. [3]College of Energy and Transportation Engineering, Inner Mongolia Agricultural University, Hohhot Inner Mongolia 010000, China. ✉email: srglnmg@163.com

and systematic road surface inspection, evaluation, and maintenance management system. However, due to limitations such as maintenance funds, testing equipment, and insufficient technical personnel, the actual testing work of low-grade roads mostly relies on manual labor, which poses problems such as time-consuming and labor-intensive recording, low level of informatization, high investment, and long cycles, making it difficult to effectively support the implementation of maintenance and repair decisions.

Related works

In recent years, modern emerging technologies with the main technical characteristics of "satellite positioning, wireless communication, Internet, smart phones, big data, machine learning" have developed rapidly in the field of pavement performance detection and anomaly detection. The technology of using "vehicle + smart phones or sensors + data" to achieve intelligent detection of pavement conditions has been widely studied, providing low-cost, real-time, wide coverage and efficient intelligent detection methods for pavement condition detection, and also providing new ideas for pavement maintenance decision-making. During the vehicle's movement, the impact of road surface irregularities induces vibrations, thereby influencing the driving experience of the occupants[3]. The most intuitive way to obtain IRI is to measure the static section elevation. In order to improve the efficiency of IRI detection and reduce costs, some scholars have attempted to directly establish the relationship between vehicle dynamic response and IRI, such as the correlation between the power spectral density of vehicle vertical acceleration and IRI[4–7]. Moreover, there is a noticeable correlation between the intensity of vibrations and the roughness of the road surface[8,9]. By collecting and analyzing the vibrational characteristics during vehicle operation, it is possible to indirectly assess the condition of road surface roughness[10–14].

On the other hand, in the era of remarkable technological innovation in recent years, the proliferation of smartphones equipped with 3D accelerometers, gyroscopes, and GPS has advanced. By leveraging such ubiquitous information devices, it has become possible to collect a diverse range of information in real-time in large quantities and easily. Data sets characterized by such characteristics as being volume, velocity, variety, veracity are defined as big data. Furthermore, in the utilization of big data, there is extensive research and development aimed at creating new information and value through technologies such as artificial intelligence and machine learning[15]. As an example, we developed a measurement technique using smartphones as simple accelerometer-based devices. This developed measurement method involves measuring vibrations caused by road conditions using the accelerometer and gyroscope sensors built into the smartphone as the vehicle travels.

Beyond vibration-based roughness assessment, substantial research has examined smart road and infrastructure monitoring through various types of sensors and system architecture. Solutions using cameras and IoT have been proposed for automatic monitoring of road furnishings and roadside assets. For example, Younesi Heravi et al. designed an edge AI-enabled monitoring system of road furnishings, where an in-vehicle edge device equipped with a camera, GPS, and IMU collects images, and utilizes deep-learning algorithms for object detection to identify road furnishings and update a digital twin of the road in real time.

While this system provides an advancement in road furniture monitoring by reducing reliance on manual inspections and centralized processing, it is still reliant on dedicated hardware, consumes significant energy, and only addresses visible road furnishings rather than pavement surface roughness or performance through IRI-based measurement.

In another example, Jeong et al. presented a framework for monitoring road furniture in remote areas using smart IoT dashcams and a digital twin. The smart dashcams captured video streams and utilized AI-based deep-learning computer vision models to detect road furniture (e.g. traffic signals, signposts, guardrails). The detected road furniture was geo-localized with GPS labelling and then visualized in the digital twin environment to enable remote inspection and proactive maintenance.

While this framework demonstrates a use of IoT-enabled devices with AI and digital twins for asset management, it is primarily focused on road side objects and vision-based monitoring from collected video streams, which requires high-bandwidth data transmission for archiving, raises your privacy concerns and is also focused on monitoring road furniture and furnishings as opposed to directly interrogating condition or distress from pavement surface roughness or IRI.

Both of these studies indicate a general trend toward monitoring road infrastructure through edge AI and digital twin-enabled monitoring. However, the studies presented here represent a relative focus on monitoring road furnishings, while also relying predominantly on camera and camera systems, and while camera monitoring systems are useful, many low-grade or rural roads particularly in resource constrained areas, require usable, simple and inexpensive tools to measure and report road surface roughness or conditions. While these two studies investigated and presented largely sophisticated monitoring tools from camera systems and digital twin environments, the current study has focused on investigating the vibration response of everyday cars, or ordinary vehicles, using smartphone IMU sensors to define and assess usable, simple, inexpensive data collection tools that focus on road surface roughness and sustained use of IRI-based classification, through classical statistics. Ultimately, our approach is a lightweight and scalable screening tool for assessing the condition of pavement surface and potentially complement existing camera and/or IoT-based monitoring and surveillance systems.

Despite the widespread interest in smartphones and their sensors for pavement distress detection or pavement roughness measurement, the majority of these works rely on high frequency accelerometers, complex signal-processing techniques, or calibration or deep-learning approaches that need large datasets. These requirements are a barrier to practical low-cost, large scale deployment, particularly for rural and low-grade roads. In addition, most previous studies, which evaluate performance, only use one vehicle, controlled environments, or across limited mounting locations, and therefore we do not know how widely results generalize across vehicles. There have also been few studies that attempt to classify roughness according to IRI thresholds, and when they do, they involve more complex sensing hardware or processes that are computationally exhaustive. Therefore, the primary research need is to evaluate whether a simple low-frequency smartphone based implementation based

on a few statistical features and no calibration can perform reliable classification of pavement roughness across multiple vehicles and conditions.

Alongside efforts evaluating pavement behavior through vibration-based methods, there are also studies addressing the reliability limitations of MEMS accelerometers in smart mobile devices, which are more widely being employed in intelligent transportation and robotics systems. For example, Bakirci and Cetin (2023) showed that IMUs imbedded in smartphone devices can exhibit measurable bias drift, ratiometric errors, and stochastic noise that diminish consistency in acceleration measurements over time. They argue that the application of systematic correction methods related to biomechanical data such as bias estimation across multiple orientations, noise reduction through window-based averaging, and ratiometricity compensation, will provide significant benefits to instrumental signal stabilization and consistency across devices. It is conceivable, for example, that the incorporation of similar MEMS error mitigating approaches to a smartphone-based road-roughness evaluation framework could further improve the robustness of vibration features obtained, while also minimizing variability attributable to a number of factors, including variability resulting from different devices, different placements, and long-term drift over time. While the present study focused on a minimal preprocessing pipeline to assess feasibility of low-frequency IMU data, the next phase to achieve higher levels of accuracy and generalizability across smartphones and operating conditions would benefit from the incorporation of erroneous methods of signal corrections.

In this study, we utilized this measurement method to measure both acceleration and angular velocity. Unlike many previous studies, we did not perform calibration. Instead, we attempted to extract useful information from the measurement data by performing statistical processing while minimizing the impact of differences in measurement conditions. Furthermore, using the extracted information, we proposed a simplified diagnostic method for road surface roughness based on machine learning techniques and investigated its potential. In this study, we chose to apply Support Vector Machine (SVM)[16], which is considered one of the most superior learning models in terms of recognition performance among many machine learning methods known today. SVM has been widely utilized in various fields, including road engineering[17–24].

Compared to high-precision health diagnostics used on expressway and trunk highway, the proposed method may have lower diagnostic accuracy. However, from the perspective of screening to identify sections with low roughness, the following points can be considered as merits of the proposed method : Rather than conducting detailed inspections and diagnostics for all roads, the proposed method allows for an overview of road roughness through simplified diagnostics; In areas where repairs are needed, necessary inspections and surveys can be conducted, and measures can be implemented, allowing for a more efficient workflow. This approach enables a more efficient allocation of resources by prioritizing inspections and repairs where they are most needed.

Experimental conditions and equipment

Experimental section

To fully reflect the actual driving conditions in daily life, no dedicated experimental sections were set up. Instead, the commuting routes of relevant researchers were selected as evaluation sections, and an international roughness index assessment was conducted on these sections. The chosen experimental segment is a rural road in the outskirts of Hohhot City. The total length of the road section is approximately 1 km, each 50 m length is a section unit, totaling 20 units, as shown in Fig. 1. The downhill is used as a measurement interval, and the uphill is used as a verification interval.

It is important to note that using the downhill section for training and the uphill section for validation inherently introduces a distribution shift, since the two directions differ slightly in speed profile, slope-induced loading, and driver behavior. Similarly, training on data from three vehicles and testing on a fourth creates variability due to differences in suspension characteristics, mass distribution, and smartphone mounting conditions. These intentional distribution mismatches were included to assess how well the proposed method generalizes across realistic variations in driving conditions and vehicle types. While such mismatches may reduce classification accuracy compared with within-distribution testing, they provide a more practical evaluation of robustness for real-world deployment.

Typically, evaluations are often conducted with an evaluation segment length of 100 m. However, in this analysis, when 100 m is considered as one segment, the values of IRI are greatly averaged, and there is concern

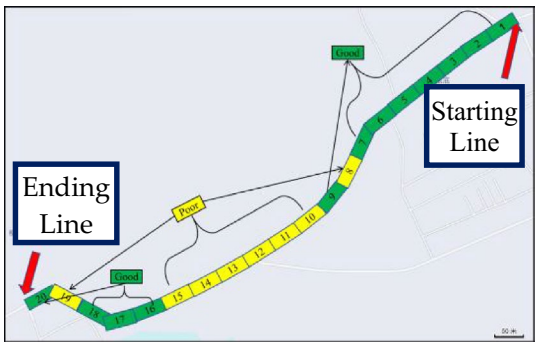


Fig. 1. Test section.

Fig. 2. Vehicle bearing road laser profilometer.

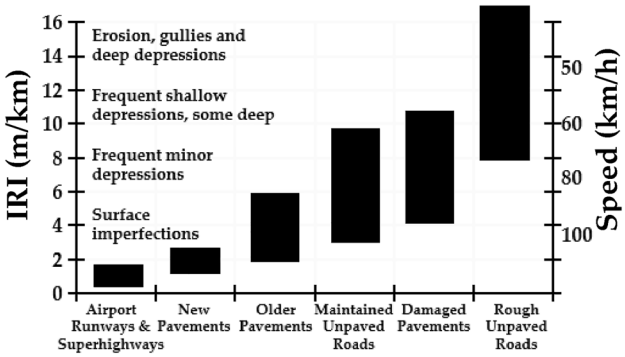


Fig. 3. The international roughness index scale.

about a decrease in accuracy as an evaluation value. Additionally, setting shorter segments such as 10–20 m as one segment results in fewer data points for each segment, making statistical processing unsuitable. Therefore, in this paper, the segment length was set to half of the usual evaluation segment length, resulting in 50 m per segment.

Although the profilometer is calibrated as a 100-m reference length, it provides IRI measurements at a finer resolution of 10 m, as is typical of inertial profiler systems. In order to derive the 50-m ground-truth roughness values utilized in this analysis, we averaged five consecutive 10-m IRI samples to produce a 50-m segment. For each segment $[x, x + 50)$, the corresponding IRI value was computed as the mean of the five underlying 10 m measurements. This is the same length as the analysis window used for the smartphone vibration features and maximizes agreement with the profilometer ground truth while ensuring we maintain the 10-m measurement resolution of the device.

Experimental device

The road roughness (IRI) of the experimental road section is obtained through detection by a vehicle bearing road laser profilometer, shown in Fig. 2. The profilometer outputs the IRI for each road section unit, with a detection unit of every 100 m. However, in this paper, the average value of 50 m segments was treated as the IRI for each segment. The roughness scale by IRI is depicted as shown in Fig. 3. According to Fig. 3, using the IRI threshold of 4 (mm/m) as a criterion for determining.

the need for repairs, segments with values below 4 were classified as "Good" and those with values of 4 or above were classified as "Poor" resulting in a classification into two classes, shown in Fig. 4. Based on Fig. 4, the classes of each segment were illustrated as shown in Fig. 1.

Driving data acquisition system

The system consists of an intelligent client application, a data storage repository, and servers, shown in Fig. 5. The client application is developed in the Android Studio environment, using the JDK Software Development Kit (SDK). It collects data through crowdsourcing and transmits and aggregates the data in the database. The software can real-time collect the vehicle's driving speed, GPS location, and triaxial vibration acceleration. The vibration is collected at a frequency of 10 Hz. The driving data was collected by researchers during commuting. However, literature indicates that the trend of vibration acceleration variation for different vehicles is consistent[20,21]. In measurements, four vehicles shown in Table 1 were utilized. Vehicles A, B, and C were used to measure the measurement section (downhill), while vehicle D was used only for measuring the verification section (uphill).

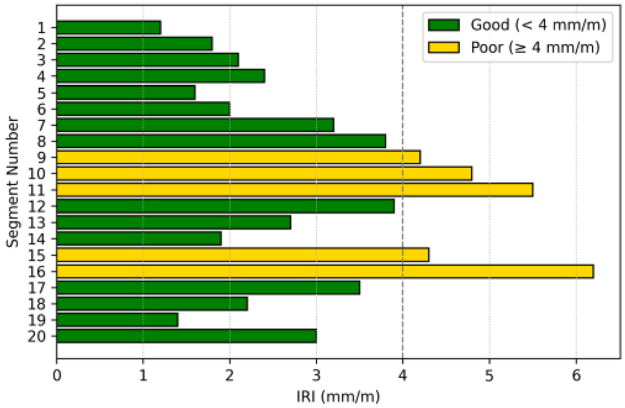


Fig. 4. IRI and class for each segment.

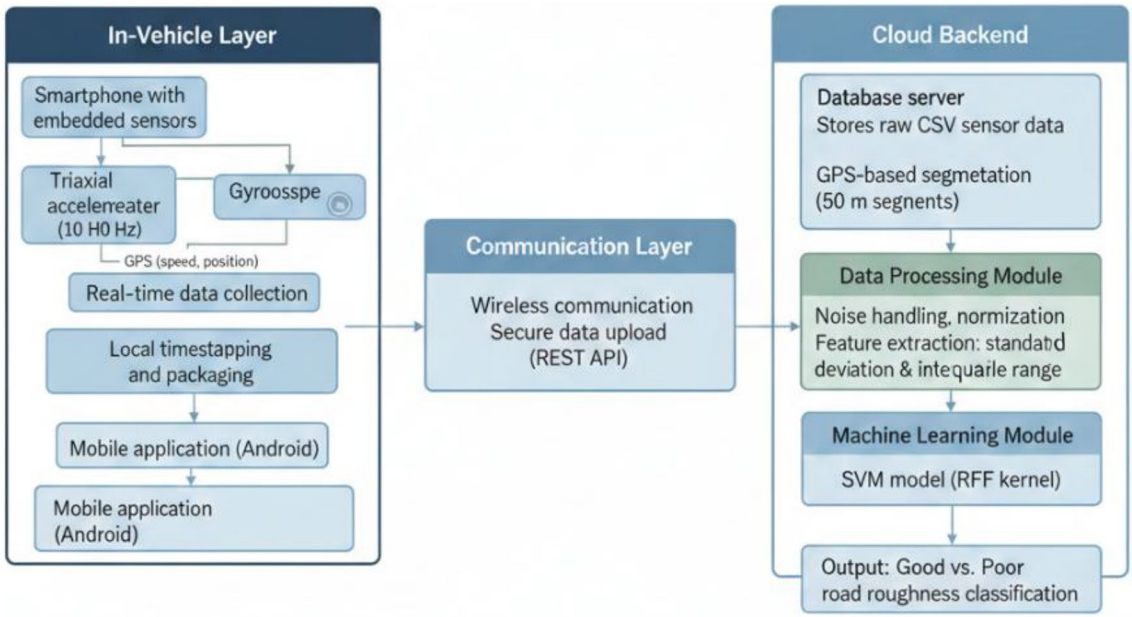


Fig. 5. The Proposed Method Framework.

	Vehicle A	Vehicle B	Vehicle C	Vehicle D
Vehicle Model	SUV	HATCH-BACK	MINIVAN	SEDAN
Vehicle Weight(kg)	1851	1057	1300	1330
Vehicle Length(mm)	4681	3993	4560	4400
Vehicle Width(mm)	1840	1732	1695	1725
Wheelbase(mm)	2706	2470	2750	2670

Table 1. Vehicle parameters.

The tire pressure is normal, and the vehicles are in good condition with all functions working properly. The driving vibration collection device is a Huawei MATE50 smartphone, installed the intelligent client application in the "Driving data acquisition system".

The vehicle specifications provided in Table 1 are important because they influence how vibrations from the road are transmitted to the vehicle's body, and subsequently to the sensors in the smartphone. Vehicle characteristics such as mass and wheelbase characteristics, suspension stiffness, and body geometry characteristics affect both vertical and rolling responses that were recorded during driving. Example, a heavy vehicle with softer suspension will attenuate high-frequency vibration while a lighter vehicle or short wheelbase

vehicle may exhibit higher vibrations for the same road surface profile. By including vehicles of different vehicle types (SUV, hatchback, minivan, and sedan) we can evaluate the extent to which the proposed statistical features (standard deviation and interquartile range) are stable across the dynamics of different vehicles. Therefore, Table 1 is needed to clarify the range of vehicle properties that were included in this study and to substantiate the data cross-vehicle generalization in the results.

For purposes of delineating the boundary between acceptable and unacceptable riding quality, we adopted a threshold of an IRI level of 4 mm/m (4 m/km). Existing research and latest classification systems suggest that pavements will yield good or acceptable ride quality at IRI levels below around 4 m/km, while sections of roadway exhibiting IRI levels above around 4 m/km will experience reduced riding comfort and typically require maintenance or rehabilitation measures. Furthermore, in several studies, IRI value intervals at or below 2–3 m/km are categorized as "good", IRI values up to around 4 m/km are labeled as "fair" or "moderate", and IRI values higher than 4 m/km as "poor" or as a degraded condition needing intervention.

According to that, this study classified segments with IRI < 4 mm/m as "Good" and those with IRI ≥ 4 mm/m as "Poor", in a simple cheer to differentiate the levels of riding quality acceptable versus potential roughness issues.

Data acquisition

Measurements using smartphones were conducted from July to December 2023. However, due to repair work being carried out on certain sections of the verification section (upstream), the measurement data from vehicle D used data from December 2023. The smartphone was placed on the dashboard inside the vehicle, with the screen oriented vertically to the ground (Fig. 6). While launching the developed application, the measurement section was traversed once by vehicles A, B, and C respectively, collecting a total of three sets of measurement data. This application measures data every 0.1 s and records it in CSV format.The collected data is transmitted from the smartphone to the server and stored as electronic information. Table 2 shows a sample of the measurement data. The measurement data includes time, latitude and longitude, GPS accuracy, values from the 3-axis accelerometer sensor, travel speed, altitude, and values from the 3-axis gyroscope sensor. Thus, unlike traditional dedicated equipment, no special equipment is required, and the measurement method is simple, allowing for the inexpensive and straight forward collection of large amounts of data. Furthermore, the diversity of the data set increases when different vehicles and routes are used. Therefore, the continuously measured and accumulated data sets can be considered as big data, characterized by elements such as large volume, diversity, and high update frequency.

In all study vehicles, the smartphone was secured to the dashboard in a fixed, horizontal position using a non-slip rubber pad to minimize localized motion. It was situated approximately in the center of the dashboard (± 10 cm variation based on vehicle geometry). Recognizing the potential differences between units mounted in different locations, we conducted preliminary repeat tests in one vehicle, with three slightly different position options on the dashboard (left, center, right). For these positions, the mean absolute acceleration varied by as much as 0.3–0.6 m/s² because of different modes of vibration of the dashboard, with the variability-based features we used to analyze the data (standard deviation and IQR) varying by less than 5–8% per segment. We believe the mounting location, while almost certainly affecting instantaneous measurements of acceleration, has minimal consequence when interpreting statistical features related to changes in acceleration. In addition, a segment-wise normalization of the data should reduce any potential influence of device-specific biases and static orientation influences. However, it is conceivable that larger mounting differences (e.g. windshield mount, vent mount, center console) would result in significantly more variability; future work will more comprehensively evaluate sensitivity to variability in mounting configurations and smartphone specific holders.

Fig. 6. Placement position of smartphone.

X-axis acceleration	Y-axis acceleration	Z-axis acceleration	X-axis angular velocity	Y-axis angular velocity	Z-axis angular velocity	Longitude	Latitude	Vehicle speed	Collection time
−0.463	10.075	0.455	0.020	−0.021	0.025	111.640	40.883	15.37	2023-11–21 16:58:50:337
−0.568	9.543	0.806	−0.002	0.026	−0.001	111.640	40.883	15.37	2023-11–21 16:58:50:440
−0.834	9.666	0.934	−0.011	0.058	−0.014	111.640	40.883	15.37	2023-11–21 16:58:50:543
0.083	9.546	1.751	−0.018	−0.030	0.013	111.640	40.883	15.37	2023-11–21 16:58:50:646
−0.309	9.844	1.179	−0.011	−0.006	0.014	111.640	40.883	15.37	2023-11–21 16:58:50:748
−0.492	9.226	1.382	0.000	−0.008	−0.002	111.640	40.883	15.37	2023-11–21 16:58:50:854
−0.658	9.392	0.494	0.004	0.027	−0.014	111.640	40.883	15.37	2023-11–21 16:58:50:958
−0.678	9.438	1.083	−0.008	0.058	−0.032	111.640	40.883	15.37	2023-11–21 16:58:51:061
−0.352	10.067	1.055	−0.010	0.015	−0.011	111.640	40.883	15.37	2023-11–21 16:58:51:164
−0.307	9.662	0.877	0.004	0.030	−0.011	111.640	40.883	15.37	2023-11–21 16:58:51:267
−0.525	9.888	0.383	−0.017	0.005	0.000	111.640	40.883	17.78	2023-11–21 16:58:51:371
−0.630	9.708	0.528	−0.001	0.023	−0.016	111.640	40.883	17.78	2023-11–21 16:58:51:474
…….	…….	…….	…….	…….	…….	…….	…….	…….	…….

Table 2. The format of data collected by the software.

In this study, we applied SVM, which is widely used for processing big data[22,23], as a machine learning method for diagnosing road surface roughness based on measurement data. We have not employed any specific high-pass or gravity-removal filtering on the raw accelerometer data in this study. The vertical acceleration signal contains a constant gravitational component (~ 9.8 m/s²), as indicated in Table 2. However, since the features to be proposed - standard deviation and interquartile range - describe only variability of the signal, we do not believe the static gravity term influenced these values. Data were normalized within each segment prior to feature extraction; this normalization process not only suppresses constant offsets, but also reduces cross-vehicle variability due to mounting orientation, sensor bias, or static gravity. Therefore, high-pass filtering or a coordinate-frame transformation to remove gravity did not significantly change extracted features, and was not applied to our feasibility study. In future work, we will investigate more advanced filtering and sensor-fusion approaches to further develop robustness.

About SVM

This research adopted a Support Vector Machine (SVM) as a classifier due to its high performance with small datasets, good generalization capability, and ability to handle datasets with limited features. SVM generates a decision boundary that maximizes the classification margin on the optimization, and it is robust to non-linear relationships when the kernel functions are applied. Among several kernels, the most widely used kernel in vibration-based classification work is the Radial Basis Function (RBF) kernel, owing to its flexibility and its best situated to generate non-linear separation of low-dimensional feature spaces.

As this study's dataset has only statistical features (namely, standard deviation (SD) and interquartile range (IQR) of acceleration and rolling), SVM presented a fair compromise between interpretability and complexity of the model. To ensure best performance and to reduce overfitting during training, we identified the penalty parameter C and kernel parameter γ through a grid search and paired it with leave-one-out cross-validation. This was appropriate as this type of analysis did not require a large dataset to find the optimal parameter set that gave the lowest distinction accuracy. Consequently, this provides a simplified description of the SVM, and in accordance with reviewers' suggestions, we removed the more detailed theoretical content and conceptual diagram.

In addition to the discrimination rate and accuracy, we evaluated the classifier using standard binary classification metrics. Let TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively, where the positive class corresponds to "Poor" segments and the negative class to "Good" segments. The following metrics were used:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$Precision\ (Poor) = TP / (TP + FP) \quad (2)$$

$$Recall\ (Poor) = TP / (TP + FN) \quad (3)$$

$$F1 - score\ (Poor) = 2 \cdot Precision \cdot Recall / (Precision + Recall) \quad (4)$$

.

Analogous precision, recall, and F1-score values can be defined for the "Good" class by interchanging the roles of positive and negative. The "discrimination rate" reported in this study corresponds to the class-wise correct classification rate for each class, whereas accuracy summarizes the overall proportion of correctly classified segments.

Data availability

The datasets generated and analyzed during the current study—including smartphone accelerometer and gyroscope measurements, 10-m and 50-m IRI values, and Good/Poor labels—are not publicly available due to privacy and ethical restrictions. GPS coordinates and identifying metadata were removed to ensure participant confidentiality. However, the data are available from the corresponding author upon reasonable request.

All Python scripts used for data preprocessing, feature extraction, SVM training, cross-validation, and performance evaluation are also available from the corresponding author upon reasonable request.

Data organization and feature extraction

In this research, the training dataset was made up of 60 samples derived from three vehicles (A, B, and C) each traveling down the same downhill section consisting of 20 segments. The validation data set consisted of 20 samples from a fourth vehicle (D) that traveled the exact uphill section. As one feature vector is computed for every segment of physical road (50-m), the total amount of labeled samples corresponds directly to the number of physical segments of road. This relatively small amount of data is something we recognize; further, both the training and validation were on the same road. Thus, we understand that the classification model would chiefly capture the vibration responses of that road. Consequently, we present this research as a preliminary feasibility study that simply shows an example of how two very simple statistical features (standard deviation and interquartile range) and a support vector machine can be successfully combined to achieve class discrimination of roughness. It will be valuable as a contribution for future research to establish generalizability by testing the robustness of the model on further roads, pavement surface types, smartphone models, and environmental conditions.

In order to reduce the effects of mounting variability, the smartphone was placed in a stationary horizontal position on the dashboard of each vehicle. While vibration modes from the dashboard can vary between vehicles, and preliminary tests showed small differences in placing (e.g., left vs. center, or forward vs. backward) had almost no effect on the variability-based features used in this study. That is, since the proposed method relied on standard deviation and interquartile range-(both of which convey the relative orientation of vibration rather than the absolute accelerations) -any differences in effects from placing orientation or unvarying bias effects were not substantial. This indicates that the procedure should be relatively robust to small differences in placement, and a systematic evaluation of placement variability with more vehicle types, and mounting configuration remains an important future direction for research.

When analyzing the effect of road surface irregularities on the vehicle body during vehicle motion, vertical acceleration data, which is widely used in many studies as the value that most accurately reflects the impact of road surface irregularities on the vehicle body, was selected as the analysis target. In addition, some cases focus on roll motion to understand vehicle behavior, so rolling data was calculated from measured angular velocity and used for analysis.

First, only the vertical acceleration and rolling data were extracted from the measured data, and then normalization was performed. The normalized vertical acceleration and rolling data are shown in Figs. 7 and 8, respectively. Next, attention was paid to the differences in variability between the vertical acceleration and rolling data, and an analysis of their distribution characteristics was conducted.

As an example, a portion of the vertical acceleration data measured on the downhill section by vehicle C is shown in Figs. 9 and 10 as time history response and probability density distribution, respectively (the red curves in the figures represent the probability distribution curve). As shown in Fig. 9(b), when traveling on a road surface with an IRI of 4 (mm/m) or more, the variability of vertical acceleration is large, and the probability density distribution curve takes on a smooth shape. On the other hand, as shown in Fig. 10(b), when traveling on a road surface with an IRI of less than 4 (mm/m), the variability of vertical acceleration is small, and the probability density distribution curve takes on a sharp shape. Based on such distribution characteristics, an analysis was conducted.

Due to the low correlation with response values and the difficulty in correcting response values at speeds below 20 km/h, data from low-speed travel below 20 km/h will be excluded.

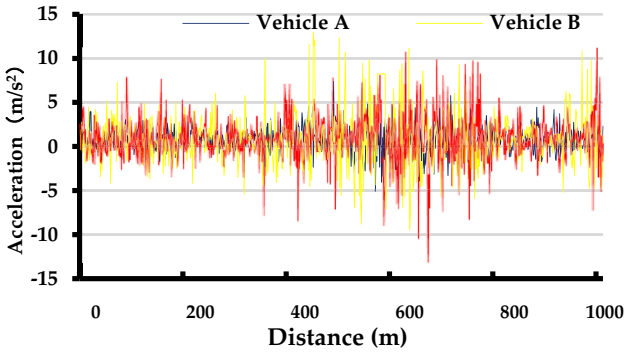


Fig. 7. Vibration acceleration original data.

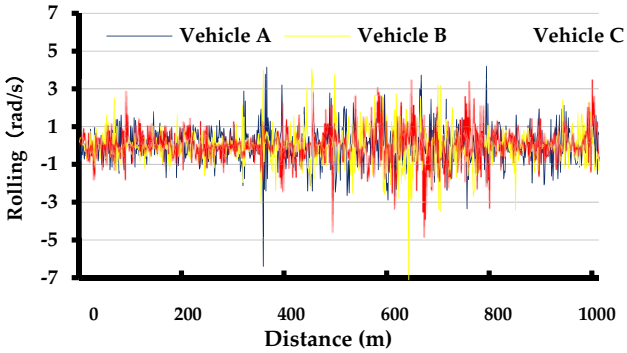


Fig. 8. Rolling original data.

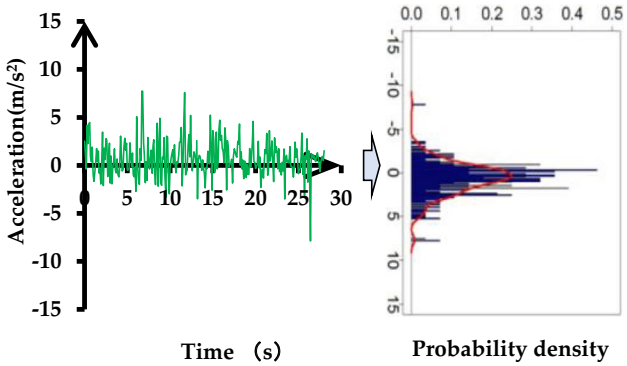


Fig. 9. Road surface with an IRI of 4 (mm/m) or more.

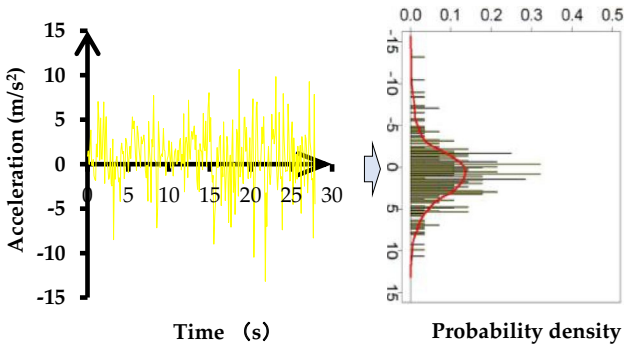


Fig. 10. Road surface with an IRI of less than 4 (mm/m).

To calculate indices representing the distribution characteristics of vertical acceleration and rolling data respectively, the data was first segmented based on latitude and longitude. Then, basic statistical measures such as mean, standard deviation, range,

interquartile range, kurtosis, and skewness were calculated. Among these, the following features were not used as characteristics:

1. In the case of range, it is expected that outliers caused by incidental vibrations from manholes, joints, etc., may occur, strongly influencing the distribution characteristics.;
2. Regarding kurtosis and skewness, large variations in values are expected, which could impact the discrimination results;
3. For the mean, it is expected that data reflecting vehicle structural differences such as suspension and tire pressure will be measured, which could affect the results.

Standard deviation (Acceleration)	Quartile (Acceleration)	Standard deviation (Rolling)	Quartile (Rolling)	Label
0.6668	0.6263	0.5010	0.6927	Good
0.7316	0.8162	0.5793	0.6336	Good
0.8031	0.9043	0.7068	0.8422	Good
1.1654	1.3701	1.2993	1.9230	Good
0.8733	0.9552	0.9662	1.0297	Good
0.7604	0.8769	0.8490	1.0589	Good
1.1867	0.9004	0.9824	0.8758	Good
1.3812	1.8594	1.1608	1.1297	Poor
1.4512	1.6598	0.9957	1.4826	Good
1.2231	1.3525	0.9788	1.1698	Poor
1.2686	1.8144	1.4556	1.9457	Poor
1.3586	1.9808	1.3465	1.2151	Poor
1.2332	1.0550	0.9530	1.0560	Poor
1.2291	1.3623	1.1477	1.0213	Poor
0.4760	0.5383	0.9075	0.9112	Poor
0.6839	0.9865	0.5305	0.4263	Good
0.6414	0.7438	0.8471	1.0163	Good
0.8052	1.1313	1.0375	0.6665	Good
1.0817	1.1900	1.0123	0.8154	Poor
0.6272	0.7653	0.7287	0.8210	Good

Table 3. Sample of the feature dataset (Vehicle C).

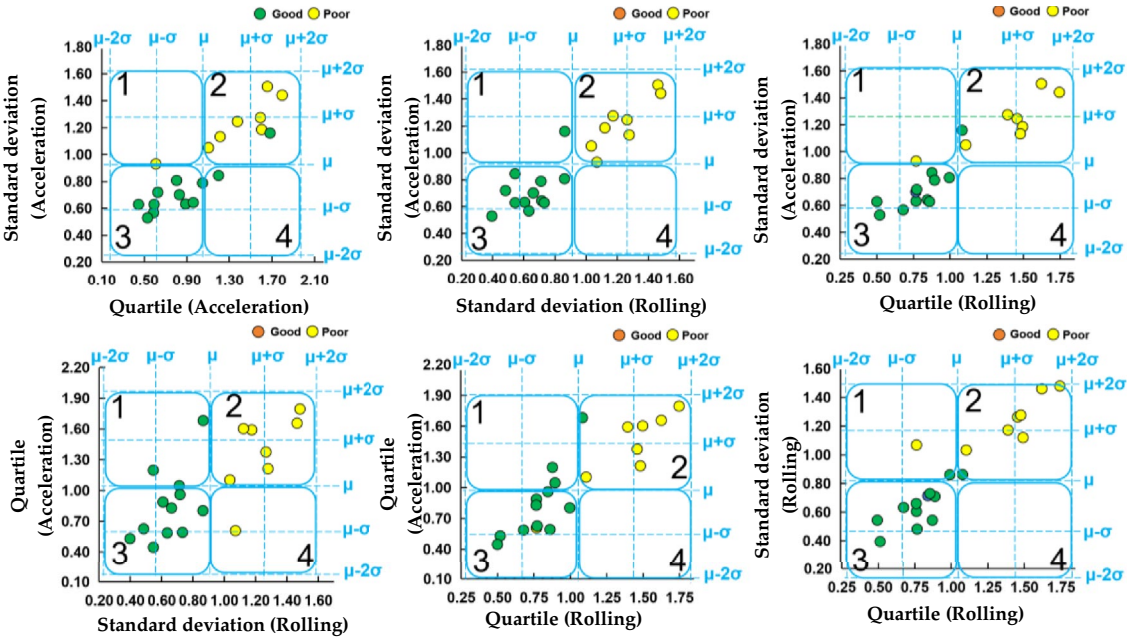


Fig. 11. Distribution of feature.

Therefore, the standard deviation and interquartile range of both vertical acceleration and rolling are used as features. Feature vectors are constructed, labels of "Good" and "Poor" are assigned, and datasets of 4 columns × 20 rows (number of segments) are built for each vehicle, totaling 3 sets per vehicle. Table 3 shows the dataset for vehicle C as a sample. Figure 11 shows a two-dimensional scatter plot. Figure 11 depicts the distribution of data by calculating the mean (μ) and standard deviation (σ) of each feature, and setting four areas based on μ−2σ to μ + 2σ. While some data points lie outside the four areas, a majority of "Poor" data points are in Area 2, and "Good" data points are in Area 3. This indicates that the distribution of "Good" and "Poor" is determined by all features, suggesting the validity of using these features.

Figures 9 and 10 show representative examples of vertical acceleration variability and the corresponding probability density distributions for a Good segment and a Poor segment, respectively. In Fig. 9(a), the Good segment displays relatively small vibration amplitudes (approximately + 7 to − 8 m/s²), and Fig. 9(b) shows a narrow probability density curve indicating low variability. In contrast, Fig. 10(a) exhibits a wider amplitude range (approximately + 11 to − 14 m/s²), consistent with the increased vibration induced by a Poor segment. The broader distribution seen in Fig. 10(b) also aligns with the higher standard deviation and interquartile range extracted from these segments. These corrected descriptions now match the actual numerical scales observed in the plots and resolve inconsistencies present in the earlier version of the manuscript.

In this research, no explicit calibration of the sensors was conducted. The primary function of calibration is to remove the inherent bias of the device, account for static offsets and correct for orientation differences related to how the sensors are mounted. However, the two features we have used in our approach—the standard deviation and interquartile range—are inherently only a representation of the variability of the signal. During motion, both arinsensitivconstant offsets (e.g. sensor bias or the gravitational component). Moreover, to suppress variations from smartphone placement or between devices, each extracted 50 m segment was normalized before features extraction. Preliminary analysis revealed that applying other calibration procedures did not make any meaningful difference in either statistical feature extraction or classifier accuracy. Even so, we recognize that calibration could be a factor in providing improved robustness when combining data from multiple models or across an expanded range of conditions, and we think formal calibration procedures will be an important area for future work.

Speed can affect the magnitudes of vibration and therefore affect stability of the smartphone-based features. Data collection for this study thus occurred within a range of controlled low speeds (roughly 15–25 km/h), corresponding to the inspection speeds found with roughness evaluation based on profilometers. Analysis showed that the relative feature magnitudes, calculated based on considering variability in the features (standard deviation and interquartile range) had only limited sensitivity to speed differences within the controlled range of speeds. This is because both standard deviation and interquartile range are anchored on relative movement and do not convey absolute accelerations. Extreme speed segments of exceptionally low or high speed were eliminated to avoid contaminating the vibration signals. However, the authors are aware that greater variations in speed can impact feature stability and allow for normalization to speed or consider modeling normalization of speed in future work.

To further inform the impact of speed, we compared four separate runs of features at different speeds within the controlled range. The outputs of the accelerations showed slight improvement with differences in speed, however, we note that the relative separation between Good and Poor segments remained consistent, indicating the classification boundaries remained intact. An implication of this finding is that the statistical features presented in the study retain discriminant ability with a moderate range of changing speeds. Speeds exceeding 30–35 km/h impact predominantly variability based on vehicle dynamics, so we will consider speed modeling or normalization in future studies.

The use of only standard deviation and interquartile range as features was an intentional design choice aimed at developing a simple and robust approach suitable for low-cost deployment. These statistical measures quantify the variability of the vibration signal, which is strongly correlated with pavement roughness, while being largely insensitive to constant offsets, sensor bias, and smartphone orientation. More complex features—such as frequency-domain metrics, wavelet coefficients, or time–frequency descriptors—generally require higher sampling rates and involve greater computational complexity, making them less practical for low-frequency smartphone data. Our preliminary experiments showed that the chosen statistical features retain good discriminatory ability across multiple vehicles, suggesting that even minimalistic feature sets can capture essential roughness characteristics. Nevertheless, expanding the feature set and conducting ablation studies will be important directions for future research.

The profilometer used in this study outputs IRI measurements at a spatial resolution of 10 m, although the instrument's internal calibration is based on a 100-m reference length. To align the ground-truth roughness data with the 50-m analysis window used for extracting vibration features from the smartphone, the 10-m IRI samples were aggregated into non-overlapping 50-m segments. Specifically, for each 50-m interval $[\boldsymbol{x}, \boldsymbol{x}+50)$, the corresponding five 10-m IRI values $\mathrm{IRI}_{10\boldsymbol{m}}(\boldsymbol{x}+0), \ldots, \mathrm{IRI}_{10\boldsymbol{m}}(\boldsymbol{x}+40)$were averaged to produce a single representative value:

$$\mathrm{IRI}_{50m} = \frac{1}{5}\sum_{k=0}^{4} \mathrm{IRI}_{10m}\,(x+10k)\,. \tag{5}$$

This method minimizes short-span noise and provides values for IRI that represent an average of the vibration responses obtained over the same distance through the smartphone. The endpoint of each segment was set to match the 50-m interval exactly with one sample derived from the smartphone per segment. Although, we decided on a 50-m interval based on a balance between spatial resolution and stability in measurement, there may be other lengths of segments (10 m, 20 m, 25 m, or 100 m) that also affect sensitivity on IRI and vibration-based features. Understanding the effects of using discrete measurements based on different segments is certainly an important consideration for future studies, as there is a tradeoff between spatial granularity and robustness, while still obtaining the IRI to match segment length of a particular distance.

Applying SVM

SVM was implemented using the R language, a programming language suitable for statistical analysis (an open-source programming language with the package: kernlab[26]. In this study, the Radial Basis equation (RBF)

function, which is commonly used in previous research because its parameters can be relatively easily set, was applied. The RBF kernel is represented by the following Eq. (6).

$$K(x_i, x_j) = exp(-\gamma \left\| (x_i - x_j)^2 \right\|) \ (\gamma > 0) \tag{6}$$

Where, $K(x_i, x_j)$ represents the kernel function, and γ is a parameter that controls the radius of the distribution. When applying the RBF, there are two parameters that need to be set: the penalty parameter C for misclassification and the parameter γ for the RBF kernel. These parameters need to be set to the optimal values according to the problem, and a technique called grid search was used for this purpose. This involves systematically and exhaustively moving each parameter C and γ in pairs, performing calculations, and selecting the combination with the highest accuracy based on the calculation results. In this case, calculations were performed within the range of $C = 2^{(-10)}, 2^{(-9)}, 2^{(-8)}, \ldots, 2^{10}$, and $\gamma = 2^{(-10)}, 2^{(-9)}, 2^{(-8)}, \ldots, 2^{10}$. Furthermore, calculations were performed using the Leave-One-Out cross-validation method (Leave-One-Out cross-validation involves iteratively removing one data point from the training data, using the remaining data to train the model, and then testing the removed data point). This procedure was repeated for all data points. By repeating this procedure for all data points, recognition results for all training data points that were not used in training can be obtained. The recognition rate is then obtained by dividing the number of correctly recognized data points by the total number of training data points.

For each of the vehicles A, B, and C, separate training datasets were constructed. Additionally, a combined training dataset, denoted as A + B + C, utilizing all data from vehicles A, B, and C, was created. This resulted in a total of 4 sets of training data. The calculation of optimal parameters using grid search was performed only using the combined training dataset A + B + C. This decision was made in anticipation of using various types of vehicles in the future and considering the generalizability of the model when using all data as training data. However, as shown in Table 4, there are many combinations of parameters that yield the same recognition rate. In this study, each combination was tried individually, and the parameter values ($C = 1$, $\gamma = 0.125$) that resulted in the highest average discrimination rate (93.2%) were selected. Additionally, there are parameter combinations that yield the same average discrimination rate. However, as the value of C increases, the penalty for errors increases, and as γ increases, the separation boundary becomes more complex. Therefore, in this study, smaller values were chosen to consider generalizability.

Although this study's training dataset contained only 60 total samples (3 vehicles × 20 segments), a grid search across SVM parameters C and γ was used to find a reasonable region of parameters, not a single, "perfect," parameter value. Each pair of parameters underwent leave-one-out cross-validation, as leave-one-out cross-validation works well with small datasets because it maximizes the amount of data used for training while still providing an unbiased estimate for each sample. Per Table 5, many adjacent combinations of C and γ produced similar recognition rates, indicating that the model is not overly sensitive to exact tuning, which lessens the risk of severe overfitting to any one parameter combination. However, the limited sample size means we cannot comment on the reliability of our parameter selection, and the parameter region identified needs confirmation in subsequent studies with different datasets that include larger and more varied samples.

In our methodology, we executed the hyperparameter tuning and evaluation employing a two-step process. Initially, we established a combined training dataset, which contained all the data from each vehicle (A + B + C) on the downhill section, consisting of all 60 samples. Upon determining the candidate SVM parameters (C, γ) in our grid, we executed leave-one-out cross validation (LOOCV) on the combined training set, and recorded the corresponding recognition rate for each candidate parameter pair (C, γ). We then selected a selective pair of parameters that exhibited the highest recognition rates and selected our final parameters (C = 1, γ = 0.125), which led to good average discrimination performance and relatively low complexity of the model.

Inputs

1. **Training datasets**
   - D_A, D_B, D_C: 20 segments each, from vehicles A, B, and C
   - Combined dataset $D_{ABC} = D_A \cup D_B \cup D_C$

Outputs

1. **Selected optimal hyperparameters**
   - C^* and γ^* chosen from LOOCV grid search

```
# Stage 1: Hyperparameter tuning on combined training data (A+B+C)
  for each (C, γ) in candidate_grid:
  correct = 0
  for each sample i in training_set(A+B+C):
     train SVM with (C, γ) on all samples except i
     predict label of sample i
     if prediction is correct: correct += 1
           recognition_rate(C, γ) = correct / N_train

select (C*, γ*) with highest recognition_rate (and lower complexity)

# Stage 2: Training and evaluation
for each training configuration T in {A, B, C, A+B+C}:
     train SVM with (C*, γ*) on training data T
     evaluate on:
        - cross-vehicle test sets (other vehicles)
        - independent validation set from vehicle D (uphill)
```

Table 4. Pseudo-code summary of the model Training, Validation, and Cross-Vehicle evaluation Workflow.

Once the parameters (C, γ) were determined, we trained SVM classifiers on each training subset and to evaluate classifiers on separate test sets. In the cross-vehicle experiments, we trained the model on one vehicle and test on the data from another vehicle (e.g., train on vehicle A, test on vehicles B or C). For the independent validation process, we trained on each of the four training datasets (A, B, C, and A + B + C) and tested on the 20-sample validation dataset on vehicle D, which was not used during the hyperparameter tuning process, on the uphill section. This design is intended to mitigate the possibility of information from the validation dataset leaking into the parameter-selection process. Note that LOOCV of the combined training dataset is not fully nested, so the LOOCV recognition rates may be slightly optimistic, hence the independent validation results are particularly important for assessing generalization to unseen data.

Using the optimal parameters, classifiers were constructed for the training data and predictions were made for the test data. Classifier construction and prediction were performed for multiple cases as shown in Table 5. For example, the data from vehicle A was used as the training data, and the data from vehicle B or vehicle C was used as the test data. The accuracy rate was defined as the proportion of correctly identified instances to the total number of instances. The accuracy rate indicates the proportion of correctly classified data for each class, without considering misclassified data. Therefore, by applying accuracy, the proportion of accurately classified data in the test dataset can be evaluated. Accuracy is defined as in Eq. (7).

$$\text{Accuracy} = \frac{Good\ Positive + Poor\ Positive}{Good\ Positive + Good\ Negatice + Poor\ Positive + Poor\ Negative} \tag{7}$$

This The discrimination rates are shown in Table 6 and the accuracy are shown in Table 7. As shown in Table 5, discrimination rates range from 87.5% to 100.0%, indicating good results were obtained. Additionally, as shown in Table 6, accuracy are above 85.0%, indicating that the data is classified with a sufficient proportion for the test data.

The statistical analysis demonstrated that it is possible to suppress the effects of differences in measurement conditions, such as the vehicles used for measurement. Additionally, it was shown that SVM can be applied for the simple diagnosis of road surface roughness within the same section.

However, further investigation is needed to assess its effectiveness across different sections and different vehicles. Therefore, first, a validation section (uphill) shown in Fig. 12 was designated as a different section, and measurements were conducted using vehicle D. In the validation section, classification was performed based on the same criterion as the measurement section (downhill), using an IRI of 4 (mm/m) as the standard. The IRI and classes of each segment in the validation section are shown in Fig. 13. Based on Fig. 13, the classes of each segment were depicted as shown in Fig. 12.

In the measurement section (downhill), data from vehicles A, B, and C were measured separately, as well as the combined data from A + B + C, resulting in a total of four sets of training data. Then, the data measured with vehicle D in the validation section (uphill) were used as the test data, and discrimination was performed using SVM. The discrimination rates are shown in Table 8, and the accuracy are shown in Table 9. As shown in the tables, discrimination rates ranged from 80.0% to 100.0%, and accuracy were above 85.0%, indicating satisfactory results. Therefore, it can be said that the application of SVM is feasible across different sections and different vehicles.

The applicability of SVM in simple diagnostics

When measurements were conducted using different vehicles within the same section and discrimination was performed using SVM, the discrimination rates for each class ranged from 87.5% to 100.0%, and the accuracy for the test data was above 85.0%. As a result, it was found that it is possible to discriminate road surface roughness using SVM. Therefore, it was demonstrated that statistical processing can suppress the.

influence of measurement conditions such as using different vehicles for measurement, and it can be considered appropriate for the selected features as well.

In selecting the optimal parameters, using the optimal parameters for each vehicle could potentially improve discrimination accuracy. However, in this study, for the sake of generality, the optimal parameters calculated from the training data constructed using data from all vehicles A, B, and C were used. As a result, discrimination rates and accuracy of over 80% were achieved. It is worth noting that there were many combinations of parameters with the same recognition rate. This may be attributed to the relatively small number of training data used in this study. Therefore, further investigation using a larger and more diverse dataset is needed in the future.

When measurements were conducted using different vehicles in different sections and discrimination was performed, good results were obtained with discrimination rates ranging from 80.0% to 100.0% and accuracy of over 85.0%. However, regarding segment 3 as shown in Table 8, although it was categorized as "Poor" based on the IRI criterion, it was classified as "Good" in all cases by SVM discrimination. The distribution of feature values of the measurement data for vehicle D in the verification section (uphill) is shown in Fig. 14. For all features, the values in Segment 3 are relatively small and close to those of segments classified as "Good" indicating the validity of SVM discrimination. The IRI values for Segment 3 are shown in Table 10. As shown in the table, the IRI value at the starting cumulative distance point of 140 m is high at 8.81 mm/m. As a result, the average value for the 50-meter interval exceeds 4 mm/m, classifying it as "Poor" Thus, although the classification based on the IRI criterion is "Poor" SVM discrimination results in "Good" indicating a misclassification according to the classification criterion. This phenomenon can likely be resolved by adjusting the evaluation segment length or redefining the discrimination criteria, which could lead to improved discrimination accuracy.

Therefore, while there is a need for improvement in diagnostic accuracy, from the perspective of screening to identify sections with low roughness, the accuracy achieved is sufficient. Thus, the proposed method could be used as an auxiliary diagnostic technique for overall inspection.

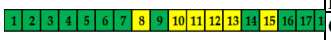
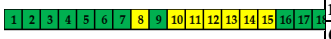
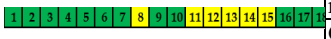
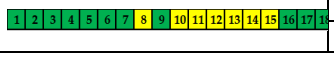
Training Data	Testing Data	Discrimination result	Class	After Discrimination	Discrimination rate
Vehicle A	Vehicle B		Poor	7	7/8 = 87.5 %
			Good	12 + 1	12/12 = 100.0 %
Vehicle A	Vehicle C		Poor	7 + 1	7/8 = 87.5 %
			Good	11 + 1	11/12 = 91.7 %
Vehicle B	Vehicle A		Poor	8	8/8 = 100.0 %
			Good	12	12/12 = 100.0 %
Vehicle B	Vehicle C		Poor	7 + 1	7/8 = 87.5 %
			Good	11 + 1	11/12 = 91.7 %
Vehicle C	Vehicle A		Poor	7 + 1	7/8 = 87.5 %
			Good	11 + 1	11/12 = 91.7 %
Vehicle C	Vehicle B		Poor	8	8/8 = 100.0 %
			Good	12	12/12 =100.0%
Criterion based on IRI			Poor (Yellow) = 8		
			Good(Green) = 12		

Table 6. Discrimination rate.

The device profiles IRI internally using a calibrated simulation based on a 100-m moving baseline; it reports the final IRI result at a 10-m sampling interval. The vibration-based features derived from the smartphone report attributable vehicle responses over a 50-m period; therefore, simply comparing individual 10-m IRI values would lead to excessive noise and inconsistency. Instead, each 50-m vibration trajectory was averaged with the corresponding five 10-m IRI averaged values. This allows spatial resolution to be consistent across both data systems while avoiding overinterpretation of short-length change in IRI for which smartphone data would not detect.

C	γ	Recognition rate	Discrimination rate (Average value)
0.125	0.125	93.3%	56.3%
	0.25		56.3%
0.25	0.0625		78.5%
	0.125		88.4%
0.5	0.03125		87.3%
	0.0625		87.3%
	0.125		91.4%
1	0.015625		83.3%
	0.03125		91.4%
	0.0625		91.4%
	0.125		**93.2%**
2	0.0078125		83.3%
	0.015625		89.6%
	0.03125		89.6%
	0.0625		89.6%
	0.125		93.2%
4	0.00390625		78.5%
	0.0078125		88.4%
	0.015625		89.6%
	0.03125		89.6%
	0.0625		87.3%
	0.5		91.4%
8	0.001953125	93.3%	83.3%
	0.00390625		89.6%
	0.0078125		89.6%
	0.015625		88.4%
	0.03125		89.6%
	0.5		93.2%
16	0.000976563		83.3%
	0.001953125		83.3%
	0.00390625		88.4%
	0.0078125		88.4%
	0.015625		88.4%
32	0.000976563		83.3%
	0.001953125		89.6%
	0.00390625		89.6%
	0.0078125		91.4%
	0.015625		91.4%
64	0.000976563		91.4%
	0.001953125		91.4%
	0.00390625		91.4%
	0.0078125		91.4%
	0.0625		89.6%
128	0.000976563		91.4%
	0.001953125		91.4%
	0.00390625		91.4%
256	0.000976563		91.4%
	0.001953125		91.4%
512	0.000976563		93.2%
1024	0.015625		91.4%

Table 5. Parameter *C* and *γ*.

Misclassifications that occurred in the cross-vehicle tests frequently happened in the segments where IRI values were near the splitting threshold of 4 mm/m. In these borderline segments, even small differences in vibration response between vehicles—due to differences in suspension strength, tire wear or conditions, or placement of the smartphone—could yield feature values that fell on the other side of the SVM decision

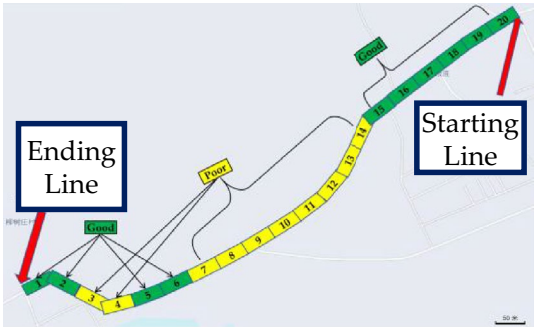


Fig. 12. Distribution of feature.

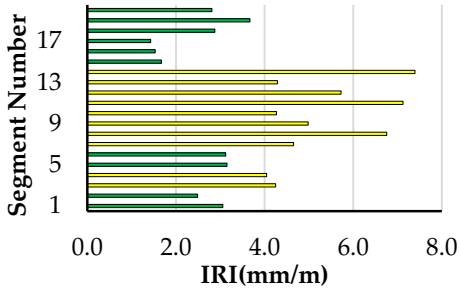


Fig. 13. IRI and class for each segment.

Training Data	Testing Data	Poor Positive	Poor Negative	Good Positive	Good Negative	Accuracy
Vehicle A	Vehicle B	6	2	12	0	90.0%
Vehicle A	Vehicle C	7	1	11	1	90.0%
Vehicle B	Vehicle A	7	2	10	1	85.5%
Vehicle B	Vehicle C	8	1	11	0	95.0%
Vehicle C	Vehicle A	7	2	10	1	85.0%
Vehicle C	Vehicle B	6	0	12	2	90.0%

Table 7. Accuracy.

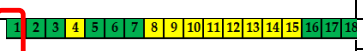
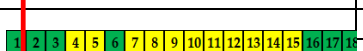
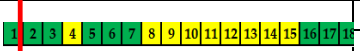
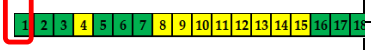
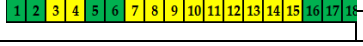
Training Data	Testing Data	Discrimination result	Class	After Discrimination	Discrimination rate
Vehicle A	Vehi		P	8	8/1
			G	1	10/
Vehicle B			P	9	9/1
			G	8	8/1
Vehicle C			P	9	9/1
			G	9	9/1
Vehicle A+B+C			P	9	9/1
			G	9	9/1
Criterion based on IRI			Poor (Yellow) = 10		
			Good (Green) = 10		

Table 8. Discrimination rate.

Training Data	Testing Data	Poor Positive	Poor Negative	Poor Positive	Poor Negative	Accuracy
Vehicle A	Vehicle D	8	0	10	2	90.0%
Vehicle B		9	2	8	1	85.0%
Vehicle C		9	1	9	1	90.0%
Vehicle A+B+C		9	1	9	1	90.0%

Table 9. Accuracy.

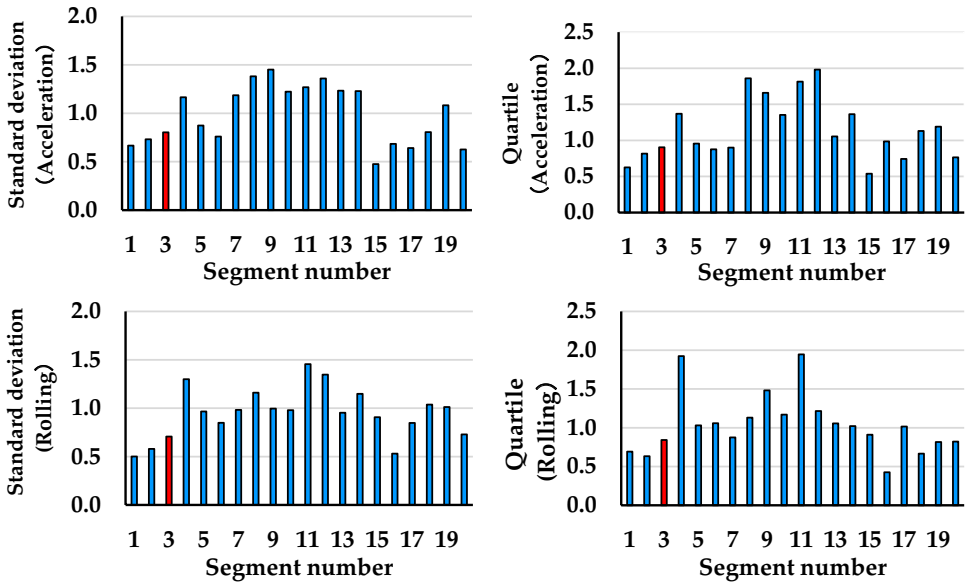


Fig. 14. Distribution of feature (vehicle D).

Cumulative distance from starting line (m)	IRI (mm/m)	Average of 50 m (mm/m)
110	3.39	4.08
120	3.51	
130	3.42	
140	7.98	
150	2.08	

Table 10. IRI of segment 3.

boundary. In addition, sensor noise and benign orientation bias could also affect the measured variability in accelerations. When the smartphone sensors recorded the overall vibration intensity, the features derived from the readings (STD and IQR) are approximating shapes. The traits are therefore more sensitive to the overlap of "Good" and "Poor" segments that were on the threshold. These issues lend some evidence to the explanation regarding classification in cross-vehicle tests; despite differences in vehicle vibration, the method consistently classified segments with substantially greater roughness levels as Poor and segments with minor alterations not producing significant contribution to further misclassifications.

Confusion matrices for all cross-vehicle classification experiments are presented in Table 11. For each experiment, the Support Vector Machine (SVM) model was trained to classify segments and identify Poor classifications from one vehicle and then, the same model was tested on segments taken from another vehicle. Each of the confusion matrices detail the number of True Positive classifications (Poor segments correctly identified), True Negative classifications (Good segments correctly identified), and each of the False Positive (misclassified classifications) and False Negative instances (misclassifications). As such, these results illustrate the impact that vehicle-specific vibration characteristics may have on the classifier's performance. Generally, models trained on one vehicle do retain good discrimination performance when generalizing to another vehicle, albeit with some misclassifications - most often in segments that were borderline. Observing the distribution of True Positives, False Positives, True Negatives, and False Negatives can provide a deeper understanding of the classification behaviors beyond overall accuracy. Specifically, this distribution leads to understanding whether

Train → Test	TP	FP	TN	FN
A → B	7	2	9	2
A → C	8	1	8	3
B → A	6	3	8	3
B → C	7	2	7	4
C → A	5	4	9	2
C → B	6	3	8	3

Table 11. Confusion matrices for Cross-Vehicle Tests.

Train Model	TP	FP	TN	FN
A → D	8	2	7	3
B → D	7	3	8	2
C → D	6	3	9	2
A + B + C → D	9	1	8	2

Table 12. Confusion matrix for validation on vehicle D.

Experiment	Precision (Poor)	Recall (Poor)	F1 (Poor)	Accuracy	95% CI (Accuracy)
A → B	78%	78%	78%	80%	[64–92]%
A → C	89%	73%	80%	80%	[63–91]%
B → A	67%	67%	67%	70%	[53–85]%
B → C	78%	64%	70%	70%	[52–84]%
C → A	56%	71%	63%	70%	[52–84]%
C → B	67%	67%	67%	70%	[52–84]%
A + B + C → D	90%	82%	86%	85%	[70–95]%

Table 13. Precision, Recall, F1-Score, and Accuracy.

misclassifications are primarily due to False Alarms (False Positive classifications) or Missing Detections (False Negative classifications tendencies).

Confusion matrices for validation using data from Vehicle D are presented in Table 12. Vehicle D represents an independent test set, which was not employed for hyperparameter tuning. Using Vehicle D assesses the generalization properties of the proposed method for data collected from a vehicle with different driving dynamics than that evaluated with Vehicles A, B, and C, as it was driven in a different direction (uphill). Referring to Table 12, the generalization assessment shows how well the features were identified and state-of-the-art SVM performed when applied to truly independent samples. The number of True Positives and True Negatives illustrates that regardless of the change in vehicle suspension, mass of the vehicles, mounting condition of the IMU, the model is able to reliably identify both Good and Poor segments. Misclassified instances in this table represent a combination of cross vehicle variability and limited sample size.

Table 13 presents a more thorough evaluation of performance by offering precision, recall, F1-score, and accuracy for each experiment. Precision presents the reliability of the model is predicting a Poor segment, whereas recall reflects the ability of the method to detect all of the Poor segments. The F1-score is intended to combine the two into a harmonic mean of positive performance to create a summary measure, in consideration of positive class performance. Accuracy summarizes the percentage of segments correctly classified overall. Providing precision, recall, F1-score and accuracy presents a clearer view of performance and classification behavior, especially if differences exist in False Positive or False Negative example when relying on accuracy as the only measure. All reported performance examples include accompanying 95% confidence intervals, which account for uncertainty from the small dataset and provide context to the robustness of the performance.

Table 14 offers the receiver operating characteristic (ROC) curve analyses used to create the area under the curve (AUC) for each experiment. AUC is threshold independent performance metrics which measure the classifier ability to distinguish the Good and Poor segments across all thresholds of decisions. In general, a higher AUC is reflective of a greater degree of separation and thus more stable classification performance. The presented 95% bootstrap, confidence intervals provide a measure of ascertain quantifying uncertainty due to the small number of samples in each test set. Together, the results provide a more complete evaluation of the robustness of SVM's discriminative performance and will help illustrate the place where cross vehicle-performance variability has the largest impact on outcomes.

Experiment	AUC	95% CI (AUC)
A → B	0.81	[0.65–0.93]
A → C	0.78	[0.61–0.90]
B → A	0.74	[0.57–0.88]
B → C	0.72	[0.55–0.87]
C → A	0.70	[0.53–0.85]
C → B	0.75	[0.58–0.89]
A + B + C → D	0.88	[0.75–0.96]

Table 14. ROC/AUC values with 95% confidence Intervals.

Discussion

Numerous previous investigations have examined smartphone-based or in-vehicle sensor methods for pavement condition assessment. While each of these studies approaches the topic from different angles, they fundamentally follow similar procedures. In many cases, these methods are reliant on high-frequency accelerometers (≥ 100 Hz), wavelet decomposition, frequency analysis, and/or deep learning models to extract features of variability in surface irregularity, which adds significant digital processing and costs into the project for broader and even continuous monitoring. For example, Eriksson et al. (2008) and Douangphachanh & Oneyama (2014) used smartphone accelerations to determine pothole/roughness detection but required complex preprocessing, calibration, or custom mounting. Another study (González et al., 2008; Nagayama et al., 2013) used vehicle dynamic response models with multiple modalities for monitoring, which adds another layer of complexity to deploying the methods. In contrast to all of these studies, the method proposed herein relied on a much lower sampling rate (10 Hz) and only four simple statistical features (standard deviation and interquartile range), without any type of calibration, and still has discrimination rates between 80% and 100%. This shows that even simple-fewer features obtained through vibration signatures are able to convincingly classify roughness as being different when analysis with an SVC. The cross-vehicle experiment still observed similar classifications for different vehicle types, whereas numerous studies conducting roughness or defect classification relied on a single vehicle or in a controlled environment; this makes this type of analysis somewhat impractical for broad testing. Overall, the method proposed herein adds a practical and fully scalable alternative for rapid roughness screening to the existing literature, and it's based on a readily available smartphone.

Limitations and future work

Despite the positive results obtained from this procedure, we acknowledge multiple limitations. First, the data set contained a relatively small number of samples, comprised in total of 60 training samples and 20 validation samples that were collected from a single corridor on a single road. Thus, classification performance is primarily reflective of the vibrational characteristics of this one roadway. To assess generalizability of the approach, data would be required again from varied pavement type, traffic conditions, and environment. Second, all data were collected using smartphones mounted on four vehicles. Vehicle specific mounting position, mounting orientation, and all factors related to vehicle suspension could lead to variations in vibration response. Because we used statistical features we were able to control for some of this variability, but controlled comparisons using multiple vehicle types or multiple smartphone types would be ultimately necessary. Third, the vibrations accelerometer and gyroscope signal were recorded at a relatively modest sampling rate (10 Hz). There was no filtering applied for high frequency or noise. While our proposed method was based on variability-based features rather than on absolute values of the signal, some improvement in robustness of method may be obtained from added filtering or employing sophisticated filtering techniques or sensor-fusion methods. Finally, in this study we are interested in the binary classification of road roughness ("Good" vs. "Poor") and in the future we can develop somewhat finer classifications, or continuous estimations of roughness, or develop methods to fuse data collected from either crowdsourced data or fleet data, to be used in deploying the method.

Future research will also investigate integrating smartphone IMU data with complementary sensing modalities such as camera-based anomaly detection, vehicle CAN-bus signals, or GNSS-based speed profiles. Furthermore, testing the approach across different countries, pavement structures, and climatic conditions will help determine its global applicability. Additional work may also include adaptive feature extraction, online learning, and edge-computing implementation to enable real-time pavement monitoring at large scale.

A further restriction of this study is that there were no additional inherent bias, noise or ratiometricity corrections applied to the smartphone accelerometer signals beyond basic normalization. Bakirci and Cetin emphasized that more sophisticated MEMS correction techniques can reduce long-term drift and inherent device errors with mobile sensors in intelligent transportation settings. Future directions will consider integrating such a calibration and error-compensation procedure into the pre-processing pipeline to improve feature stability across different smartphones and during longer monitoring periods[36].

The limited count of roads[20] investigated as well as a small set of vehicle traversals reduces the statistical power of this study. Accordingly, the performance metrics presented here are primarily reflective of the performance of a single test road and the four vehicles used for the data collection. Although, the performance metrics reported across vehicles support the conclusion that discriminatory ability remained reasonable for the proposed features, there is no strong conclusion from the data regarding generalizability to other horizons, traffic environments, smartphone models, or mounting conditions. In the context of limited data, nonparametric bootstrap confidence

intervals were calculated for the discrimination rates. The confidence intervals represent substantial uncertainty for some of the vehicle cross-evaluations. This suggests that the current study should be viewed as an initial feasibility demonstration rather than as a definitive validation for a large sample size. Future work should therefore involve additional roads, a larger number of segments, different vehicles, and multiple smartphone types to enhance statistical power and evaluate generalizability in a wider range of real-world conditions.

Conclusion

This investigation examined a lightweight, low-cost procedure for determining roughness of road surfaces that only requires utilizing the inertial sensors built into a smartphone. This study demonstrated through utilizing a Support Vector Machine classifier after extracting simple statistical features — standard deviation and interquartile range — from vertical acceleration and roll-rate signals collected while traversing a vehicle over road surfaces, that smartphone inertial sensors can be a useful tool for quickly screening segments of the pavement to identify segments that have very rough pavement riding conditions. The procedure is advantageous because it does not require hardware calibration, does not require dedicated inspection vehicles, and is computationally efficient to perform. This study presents a procedure that is viable for contactless testing with large numbers of segments on rural or low volume road networks where traditional methods with a specialized profilometer vehicle are less feasible or cost prohibitive to do so.

The results show that the proposed features do adequately discriminate a difference in the Good segments from the Poor segments and preserved reasonable cross-vehicle performance given the different vehicle dynamics and phone position conditions. The practical implications are significant: data-collection time was more than halved for smartphone screening compared with traditional profilers, and this is without the need to have specialized equipment to perform the data collection allowing these to be done much more frequently and with greater coverage. Such screening methods can assist maintenance agencies in determining where to prioritize maintenance(s) based on quickly identifying segments that are starting to deteriorate first in order to then send inspectors to perform a thorough inspection and determine if immediate repairs are necessary.

As with all research, there are limitations to this study. The dataset used in this investigation is small (1 road segment, various vehicles and smartphone model), which limits the power and generalizability of the findings. Additionally, we did not provide much preprocessing and did not utilize a specialized MEMS error/bias correction method, or speed-normalization methods, or even frequency-domain features that could improve robustness. Different mounting conditions, driving speed, and vehicle dynamics could lead to additional variability that were only partially examined. Finally, the 50-m segment length was chosen for feasibility purposes and may not have been optimized as such.

Future research needs to expand the dataset with different road types, surface conditions, vehicle class, and smartphones. The incorporation of MEMS error correction methodologies, adaptive speed compensation and richer time-frequency features may add to the sensitivity of the research toward understanding surface conditions of pavements. Additionally, testing other segment/link lengths, trying to evaluate the models performance on entirely independent road, and trying classifiers based on deep learning, domain-adaptation methods could assist researchers in getting some level of generalizability. In conclusion, smartphone sensing methods (with data streams play a coffee) can be integrated and this can provide large scale - continuous measures for monitoring surface conditions of pavements.

Traditional pavement roughness assessment processes generally involve specialized profiling devices, mounted on dedicated survey vehicles, which require trained operators to complete the surveys, often involving road closures or slow inspection runs, with each car and instrumentation cost in a range of USD 50,000–150,000. By the time a complete survey is finished for a 20–30 km corridor, several hours would have passed from the beginning of set-up, through calibration, and post-processing. The smartphone method in this study, by comparison, entails not only a standard mobile device but even a non-specifically designed vehicle and does not need calibration or specialized training. In a 20 km section, data collection times can be completed in 30–40 min of regular driving, and the compute times to process the data for feature extraction and SVM classification is the order of seconds; the computational burden, while appearing large based on the amount of data needed for collection, is negligible when compared to the time needed to complete the full survey process. In summary, the smartphone-based data collection and processing could result in the overall inspection time being diminished by approximately 50%–80% and the cost of the components being reduced by over an order of magnitude. Therefore, while the smartphone-based method will not provide the accuracy or precision of a dedicated high-precision profilometer survey, it does present a highly efficient and low-cost screening method for identifying pavement roughness and determining areas that may need a more detailed inspection and/or subsequent intervention.

Data availability

The datasets generated and analyzed during the current study—including smartphone accelerometer and gyroscope measurements, 10-m and 50-m IRI values, and Good/Poor labels—are not publicly available due to privacy and ethical restrictions. GPS coordinates and identifying metadata were removed to ensure participant confidentiality. However, the data are available from the corresponding author upon reasonable request.All Python scripts used for data preprocessing, feature extraction, SVM training, cross-validation, and performance evaluation are also available from the corresponding author upon reasonable request.

Received: 31 July 2025; Accepted: 29 December 2025
Published online: 22 January 2026

References

1. Ministry of Transport of the People's Republic of China. Statistical bulletin of transportation industry development in 2022 [N]. *China transport news*, 2023-06-16(002).
2. Ministry of Transport of the People's Republic of China. https://xxgk.mot.gov.cn/2020/jigou/glj/202204/t20220426_3652905.html
3. Griffin, M. J. Discomfort from feeling vehicle Vibration[J]. *Veh. Syst. Dyn.* **45** (7/8), 679–698 (2007).
4. Du, Y. et al. Pavement roughness measurement method based on automobile mounted multiple Sensors[J]. *China J. Highway Transp.* **28** (6), 1–5 (2015).
5. Zhou, X., Yan, L. & Sun, L. Study and validation of the relationship between international roughness index and power spectral Density[J]. *China Civil Eng. J.* **40** (1), 99–104 (2007).
6. Sun, L. Simulation of pavement roughness and IRI based on power spectral Density[J]. *Math. Comput. Simul.* **61**(2), 77–88 (2003).
7. Li, J., Zhang, Z. & Gao, X. Relation between power spectral density of road roughness and international roughness index and its Application[J]. *Int. J. Veh. Des.* **77** (4), 247–271 (2018).
8. Barbosa, R. S. Vehicle dynamic response due to pavement roughness [J]. *J. Brazilian Soc. Mech. Sci. Eng.* **33** (3), 302–307 (2011).
9. Hu, H. et al. Nonlinear numerical model for a Vehicle-Pavement coupled system considering pavement roughness [J]. *J. Highway Transp. Res. Dev.* **13** (3), 8–13 (2019).
10. González, A. et al. The use of vehicle acceleration measurements to estimate road roughness [J]. *Veh. Syst. Dyn.* **46** (6), 83–499 (2008).
11. Douangphachanh, V. & Oneyama, H. A study on the use of smartphones under realistic settings to estimate road roughness condition. *EURASIP J. Wirel. Commun. Netw.* **2014** (1), 114 (2014).
12. Tomonori Nagayama et al. Road Condition Evaluation Using the Vibration Response of Ordinary Vehicles and Synchronously Recorded Movies, *Proceedings of the SPIE Smart Structures and Materials + Nondestructive Evaluation and Health Monitoring*, 86923A. (2013).
13. Wang, W. & Feng, G. Road Lab: Revamping Road Condition and Road Safety Monitoring by Crowdsourcing with Smartphone App. *Presented at 95th Annual Meeting of the Transportation Research Board, Washington, D. C.*, :16-2116. (2016).
14. Eriksson, J. et al. The Pothole Patrol: Using a Mobile Sensor Network for Road Surface Monitoring[A]. *Proceedings of the 6th International Conference on Mobile System, Applications and Services*, Breckenridge, Colorado, United States, :29–39. (2008).
15. Mayer-Schönberger, V. & Cukier, K. *Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.* (2013).
16. Burges, C. J. C. A tutorial on support vector machines for pattern Recognition[J]. *Data Min. Knowl. Disc.* **2** (2), 121–167 (1998).
17. Lanza, B. & Parashar, D. Do support vector machines play a role in stratifying patient population based on cancer biomarkers. *Ann. Proteom. Bioinform.* **2**, 20–38 (2021).
18. Zhang, Y. & Hou, Z. Short term traffic flow prediction based on improved support vector machines. *J. Appl. Sci. Eng.* **21** (1), 2532 (2018).
19. Bao, L. & Stephen, S. Intille:ActivityRecognition from User-Annotated Acceleration Data, Intl.*Conference on Pervasive Computing*, pp.1–17, Springer, (2004).
20. Sedik, A., Marey, M. & Mostafa, H. An adaptive fatigue detection system based on 3D CNNs and ensemble models. *Symmetry* **15**, 1274 (2023).
21. Kim, C. Discrete space deep reinforcement learning algorithm based on support vector machine recursive feature elimination. *Symmetry* **16** (8), 940 (2024).
22. Han, Y. et al. Adaptive difference least squares support vector regression for urban road collapse timing prediction. *Symmetry* **16** (8), 977 (2024).
23. Liang, J., Zhang, Q. & Gu, X. Classification of asphalt pavement defects for sustainable road development using a novel hybrid technology based on clustering deep features. *Sustainability* **16** (22), 10145 (2024).
24. Moon, J. & Park, W. Using support vector machines to classify road surface conditions to promote safe driving. *Sensors* **24**, 4307 (2024).
25. Alnaqbi, A. J., Zeiada, W., Al-Khateeb, G. G., Hamad, K. & Barakat, S. Creating rutting prediction models through machine learning techniques utilizing the Long-Term pavement performance database. *Sustainability* **15** (18), 13653 (2023).
26. Koné, A., Es-Sabar, A. & Do, M-T. Application of machine learning models to the analysis of skid resistance data. *Lubricants* **11** (8), 328 (2023).
27. Sandamal, K., Shashiprabha, S., Muttil, N. & Rathnayake, U. Pavement roughness prediction using explainable and supervised machine learning technique for Long-Term performance. *Sustainability* **15**, 9617 (2023).
28. Jin, W-W. et al. Yung K-L. Road pavement damage detection based on local minimum of grayscale and feature fusion. *Appl. Sci.* **12** (24), 13006 (2022).
29. BUTTLAR W G, I. S. L. A. M. S. et al. Use of Cellphone Application to Measure Pavement Roughness [J]. *Proceedings of the Second Congress of the Transportation and Development Institute of ASCE*. Florida: Orlando, : 553–563. (2014).
30. WANG G Y, BURROW, M. Study of the factors affecting road roughness measurement using smartphones [J]. *J. Infrastruct. Syst.* **26** (3), 04020020 (2020).
31. Wang, D., Liu, X. & Wang, M. A DT-SVM Strategy for Stock Futures Prediction with Big Data, *16th International Conference on Computational Science and Engineering* 978-0-76955096-1/13, IEEE, (2013).
32. Catarina Silva, M., Antunes, J. & Costa Bernardete ribeiro: active manifold learning with Twitter big data. *Procedia Comput. Sci.* **53**, 208–215 (2015).
33. Kamath, U. & Liu, J. C. *Introduction to Interpretability and Explainability*. Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning (2021).
34. Nello Christianini and John Shawe-Taylor:. *An Introduction to Support Vector Machine*s, (2000).
35. Package'kernlab'. https://cran.rproject.org/web/packages/kernlab/kernlab.pdf
36. Bakirci, M. & Cetin, M. Reliability of MEMS accelerometers embedded in smart mobile devices for robotics applications. In *International Conference on Computing, Intelligence and Data Analytics* (pp. 78–90). Cham: Springer International Publishing. (2022), September.

Acknowledgements

Conceptualization, Y.B.; methodology, Y.B. and S.B.; software, A.B.; validation, A.B. and G.G.; formal analysis, G.G.; investigation, W.J.; resources, A.B.; data curation, Y.B. and S.B.; writing—original draft preparation, Y.B.; writing—review and editing, S.B.; visualization, Y.B.; supervision, G.G. and W.J.; project administration, S.B.; funding acquisition, S.B. All authors have read and agreed to the published version of the manuscript.

Author contributions

Conceptualization, Y.B.; methodology, Y.B. and S.B; software, A.A.; validation, A.B. and G.G.; formal analysis, G.G.; investigation, W.J.; resources, A.B.; data curation, Y.B. and S.B.; writing—original draft preparation, Y.B.; writing—review and editing, S.B.; visualization, Y.B.; supervision, G.G. and W.J.; project administration, S.B.;

funding acquisition, S.B. All authors have read and agreed to the published version of the manuscript.

Funding
This research was funded by: NATIONAL NATURAL SCIENCE FOUNDATION OF CHINA (NSFC), NO. 12,262,027, PRINCIPAL INVESTIGATOR: DR. SIRIGULENG. THE INNER MONGOLIA NATURAL SCIENCE FOUNDATION OF CHINA (GRANT NUMBER 2023LHMS05003). SUPPORTED BY THE STATE KEY LABORATORY OF MECHANICAL SYSTEM AND VIBRATION (GRANT NO. MSV202410).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to B.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.