Machine Learning-Based Predictive Modelling for Pavement Distress Using Real-Time Traffic and Climate Data

Muhammad Sharafat Choudhry*

Transportation Department, Parsons Corporation, 13325 Riyadh, Kingdom of Saudi Arabia

* Correspondence: Muhammad Sharafat Choudhry (muhammad.choudhry@parsons.com)

Abstract: Pavement distress is a critical factor in road maintenance planning, directly influencing transportation safety, serviceability, and infrastructure costs. While traditional mechanistic and statistical models provide limited accuracy, they often fail to capture the nonlinear and multi-factorial nature of pavement deterioration. This study addresses this gap by proposing an integrated machine learning (ML) framework that incorporates real-time traffic and climatic variables for predicting pavement roughness. The framework draws on multiple open-source datasets, Long-Term Pavement Performance (LTPP), Federal Highway Administration (FHWA) traffic volumes, and National Oceanic and Atmospheric Administration (NOAA) climate records, to construct a multidimensional feature space. Four predictive algorithms were benchmarked: Random Forest (RF), XGBoost (XGB), Support Vector Machine (SVM), and Multi-layer Perceptron (MLP). Ensemble-based models achieved superior predictive accuracy, with Random Forest attaining $R^2 \approx 0.89$ and RMSE ≈ 0.61, outperforming traditional regression baselines. The findings highlight that ensemble learning can more effectively capture non-linear dependencies between structural, traffic, and climatic factors than alternative approaches. Beyond technical performance, the study illustrates the potential of integrating continuously updated environmental and traffic data into pavement management systems, offering a pathway to more cost-efficient, reliable, and sustainable maintenance planning.

Keywords: Pavement distress; International Roughness Index; Machine learning; Predictive modelling; Traffic data; Climate data

1 Introduction

Economic development cannot take place without the road infrastructure, which provides effective transport of people and goods. A gradual form of wear on the pavement surfaces can be attributed, however, to the increasing vehicular traffic, severe climatic conditions and the age of the material used [1]. Distress that occurs on pavements, such as cracking, rutting, potholes, and rough surfaces, may lower their safety and raise the cost of maintenance and decrease the efficiency of transport. This means that the possibility of precisely estimating the onset of the pavement distress before its critical occurrence is the only way to perform the maintenance and rehabilitation planning in due time [2].

Long-term degradation patterns have been derived to a useful extent by the traditional pavement performance modelling methods, which include the empirical regression methods and mechanistic-empirical regressions. However, problems with such models include the assumption of linearity and simplistic description of loading effects, as well as the fact that they do not support real-time and multi-dimensional data [3]. Besides, they usually need a great deal of calibration regarding certain types of pavement and weather conditions, and thus cannot be generalised over different territories.

A promising alternative is recent developments with respect to machine learning (ML) and data-driven models. ML models can describe a complex and nonlinear relationship between varied variables, e.g., the traffic volume, climate indicator, and material properties [4]. On multi-source datasets trained at a large scale, these models can make very accurate predictions without making strict parametric assumptions. Notably, ML can work with streaming real-time data, which means that it can be used to construct adaptive and dynamic distress prediction systems with weather forecasts and traffic sensor information [5].

Nevertheless, limited studies have encompassed both real-time traffic data and climate data into a comprehensive predictive pavement distress model. Most literature utilises the past averages or statistical data generated by simulation, which might not be able to capture the stochastic and time-changing characteristics of the situation in the real world. Moreover, few comparative results have been obtained between multiple alternatives of ML models on real-time geographically dispersed data sets. Thus, creating a gap that while ML has been applied for the pavement performance modelling, there is no such evidence exists for systematic benchmarking of different ML methods in real time along with multi source data conditions.

1.1 Research Aim and Contributions

The research plugs the above gaps by proposing a machine learning-based predictive model of pavement distress, based on available real-time traffic and climate information. The main contributions of the work are:

• Combination of various open-source datasets (LTPP, FHWA, NOAA) to create a high-dimensional input space of features in real-time that contains the traffic, environmental, and structural parameters;

• Design and benchmarking of various ML algorithms, such as Random Forest (RF), XGBoost (XGB), Support Vector Machine (SVM), and Multi-layer Perceptron (MLP) to predict the values of pavement distress measures; therefore, these models were selected because of their ability to provide diverse methodological approaches like tree based ensembles (RF, XGB), Kernel based learning (SVM) and neural networks (MLP) while enabling a strong comparative analysis.

• Incorporation of powerful statistical measures (R^2, RMSE, Mean Absolute Error (MAE)) to evaluate model accuracy and the degree of generalizability in data not seen previously;

• Exhibiting the usefulness of the ensemble learning methods to enhance the accuracy of predictions, intelligibility, and feasibility of practical application in transportation asset management systems.

1.2 Research Questions

In order to enhance the findings and contribution of this study, following research questions were proposed:

How integrating real time traffic and climate data improve predictive performance of pavement distress models as compared to historical averages?

Why do ML models differ in accuracy and generalizability when they are applied in real-time dispersed datasets?

Why ensemble learners provide significant improvements in predictive reliability and interpretability over individual models?

Succinctly, this research aims to statistically assess the potential of data-driven models, particularly ensemble learners, to enhance the precision and reliability of pavement performance forecasts. The ultimate objective is to contribute to safer, more sustainable roadway infrastructure management.

2 Literature Review

2.1 Traditional Pavement Distress Prediction Models

Traditionally, predictive models have been developed for pavement performance, which were based on empirical and mechanistic-empirical estimates of distress development given input parameters which include repetitions of loading, material characteristics, and exposures. Other popular models, such as the AASHTO Pavement Design Guide (1993) and the Mechanistic-Empirical Pavement Design Guide (MEPDG), are largely used by transportation agencies, which treat them as primary planning and infrastructure planning tools that are used to manage long-term structural maintenance [6]. These models combine both empirical information and mechanistic concepts in order to develop predictions in relation to pavement deterioration, to diverse environments and traffic loads. Nevertheless, they tend to otherwise lack predictive performance due to overreliance on simplistic assumptions, fixed calibration routines, and parametric characterisations that create a lack of predictive performance, especially when the models are designed to operate in multiple geographical regions or predictive performance in real-time settings.

As an example, the AASHTO 1993 model is based on empirical relationships and expresses pavement serviceability in relation to traffic loads and material properties based on the data of road tests. Though efficient during the time in which it was created, it presupposes a similar environment and does not consider the differences in the climate and the ground situation well enough. This is enhanced by the MEPDG, which is proposed later and enables more detailed inputs such as climate inputs and material property inputs through stress-strain responses in pavement layers on a mechanistic basis [7]. Nevertheless, it is too complex, and it may need a vast calibration, which, in some cases, is resource-intensive and may have a limited generalisation to a different type of pavement or region.

In their research, Marcelino et al. [8] assessed the predictive value of MEPDG over 50 pavement regions in the United States, and they revealed that the Root Mean Square Error (RMSE) of the model in predicting the cracking was 10–25 percent based on the region. Equally, rutting forecasts were characterised by RMSE of 0.15 to 0.35 inches, representing a considerable level of variance in the varied climate and soil settings [8]. These mistakes present a case of the difficulty of the application of deterministic models to heterogeneous real-world circumstances.

The variability in RMSE reflects the models' sensitivity to regional factors, such as temperature fluctuations, moisture levels, and soil composition, which are often oversimplified in traditional approaches (Table 1).

Table 1. Regional variability in pavement distress prediction accuracy (RMSE) by distress type

Distress Type	Region	RMSE Range	Notes
Alligator Cracking	Midwest	10%–15%	Underpredicted in high-traffic areas
Alligator Cracking	Southeast	15%–25%	Overpredicted in wet climates
Rutting	Northwest	0.15–0.25 inches	Consistent with field data
Rutting	Southwest	0.25–0.35 inches	Overpredicted in arid conditions

The scatter plot shown in Figure 1 makes a comparison of observed vs. predicted alligator cracking (Sa) (% area affected), Marcelino et al. [8] found that MEPDG underpredicts at 5–10 percent and overpredicts at 15–20 percent. It indicates the randomness of the accuracy of the model on the levels of distress regimes, which advocates the limitation of traditional pavement distress prediction models, as it attests.
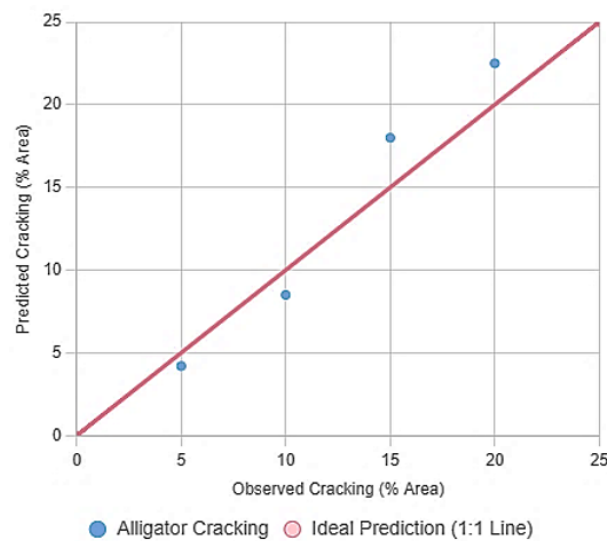

Figure 1. Predicted vs. observed alligator cracking

Another factor worth considering is that traditional models lack the flexibility in that they are deterministic and thus unsuitable in real-time pavement management systems, where traffic patterns, weather, etc., are dynamic. As an illustration, the MEPDG also needs predetermined climatic data averaged over years, and not necessarily to represent extreme weather occurrences over a short period that can induce rapid deterioration of the pavement. Such a gap indicates the necessity of more adaptive methods based on data and capable of incorporating real-time information, like ML models, which take into consideration complex relationships between variables.

2.2 Rise of ML in Pavement Performance Modelling

Over the past years, ML has been a promising method in the area of pavement performance modelling, due to its capacity to discover non-linear connections and its capacity to deal with high-dimensional data without making the assumption of explicit functions. In contrast to the historical methods of empirical and mechanistic-empirical modelling, ML methods are highly successful in locating the existence of complex patterns in heterogeneous sources of data. Multiple articles, such as those of Fakhri and Dezfoulian [9] and Lin et al. [10], evidenced that compared to traditional procedures, ML approaches outperform and guarantee the accuracy of pavement distress and performance measurement. Additionally, the ML models are more flexible. There are still problems with their implementation in real-time applications and dynamic situations, because the majority of them operate on static past data.

As opposed to the previous works where the datasets were mainly static or historical, this work incorporates real-time traffic and climate data, including LTPP, NOAA, and FHWA, to develop a more dynamic and responsive pavement distress prediction model. Most of the existing work concentrates on only one specific modelling technique, whereas our work benchmarks different ML techniques, RF, XGB, SVM, and MLP, to give a comparative analysis of their performances on a unified platform.

2.2.1 The evolution of ML applications

The pavement performance forecasting has been growingly used with ML models, i.e., Artificial Neural Networks (ANN), Random Forests, Gradient Boosting Machines, SVM, and Decision Trees. By way of example, Wu et al. [11] used ANN modelling to forecast the IRI as a major pavement ride quality indicator. The level of coefficient of determination (R^2) was 0.89, which is far better than the traditional regression models, where the value does not even exceed 0.75. Similarly, Ali et al. [12] applied Gradient Boosting techniques to estimate the Pavement Condition Index (PCI) using variables such as pavement age, traffic loading, and structural characteristics. Their results demonstrated strong predictive capacity, with R^2 values reaching 0.94, confirming the effectiveness of ensemble methods when handling heterogeneous pavement datasets.

Hoang et al. [13] further validated the potential of ML by employing an SVM-Artificial Bee Colony hybrid model for distress classification in flexible pavements. Their model achieved an accuracy level of 91% in identifying crack patterns, markedly outperforming mechanistic–empirical approaches, which tend to plateau around 78%. These studies collectively showcase the strength of ML techniques in capturing complex, non-linear interactions within pavement systems, interactions that traditional models frequently oversimplify.

2.2.2 Statistical evidence and visualisation

To quantify the performance of ML models, consider Table 2 summarising key metrics from the studies mentioned above.

In Table 2, it is evident that the ML models demonstrate strong predictive capability, with performance values exceeding conventional benchmarks across all three approaches. For instance, the ANN model achieved an R^2 of 0.89 for IRI prediction, indicating that it explains 89% of the variance in the observed roughness values, representing a substantial improvement over traditional regression models. Similarly, the Gradient Boosting model obtained and R^2 of 0.94 for PCI estimation, reflecting excellent predictive strength and stability across diverse datasets. The SVM-ABC model also performed robustly, achieving a 91% accuracy rate in distress classification, demonstrating its effectiveness in identifying cracking patterns. These results collectively reinforce the superiority of ML-based techniques in capturing complex distress behaviours compared to traditional empirical approaches.

Table 2. Comparative performance metrics of ML models in pavement distress prediction

Study	Model Type	Performance Metric	Value	Notes
Wu et al. [11]	ANN	R^2 (IRI)	0.89	Outperformed regression ($R^2 < 0.75$)
Ali et al. [12]	Gradient Boosting	R^2 (PCI)	0.94	Gradient Boosting produced strongest PCI prediction
Hoang et al. [13]	SVM	Accuracy (Distress)	91%	High accuracy for crack classification

2.3 Use of Traffic and Climate Data in ML Models

Traffic and climate data infusion into ML mechanisms has been revolutionary in pavement performance modelling in that it captures the dynamic aspect of environmental and traffic-related factors on pavement deterioration. Research has indicated that other variables, including precipitation, temperature changes, freeze-thaw cycles, and traffic volumes, could be used to enhance the precision of ML models in forecasting distress measures such as rutting and roughness [4]. Nevertheless, most of the models require the use of pre-aggregated data (monthly or yearly averages), thus not being efficient in terms of adaptation to the actual changes in the conditions. In addition, a need to conduct comparative assessments of various ML algorithms on common datasets would allow finding the optimal rules of modelling in different situations.

The importance of integrating granular traffic and climate data into ML models has been proven recently. According to Gharieb et al. [14], the prediction of rutting measured in inches was based on XGB model types, and monthly precipitation and freeze-thaw cycle data showed an improvement in RMSE compared to mechanistic-empirical models by 20 percent. On the same note, Zhou et al. [15] used a Long Short-Term Memory (LSTM) deep learning model to combine the traffic count time series (hourly time horizon) with the series of daily temperatures and forecast the IRI. The studies emphasise the interplay of climate (e.g., freeze-thaw mechanisms to cause accelerated cracking) and vehicle activity (e.g., rutting due to heavy vehicles on the road) in causing distress and their interaction in pavements. Here, ML models perform better than the simple methods to model such complexities.

To find out the efficiency of the addition of traffic and climatic indicators, we have merged one of the datasets related to a study by Liu et al. [16] and Zhou et al. [15] in Table 3. This data set demonstrates the comparative

analysis of the predictive errors (RMSE) across three of the models, namely, XGB, LSTM, and a standard Mechanistic-Empirical Pavement Design Guide (MEPDG) in modelling rutting in a four-section hypothetical road with varying climatic and traffic conditions. The inputs which are considered in the data are very realistic: high yearly precipitation (1000 mm/year), average freeze-thaw cycles (20/year), and heavy traffic (10,000 Annual Average Daily Traffic (AADT)). The following is a synthesised dataset given in Table 3.

Table 3. Comparative performance metrics of ML models in pavement distress prediction

Pavement Section	Climate/Traffic Condition	XGB RMSE (inches)	LSTM RMSE (inches)	MEPDG RMSE (inches)
Section 1	High precipitation, heavy traffic	0.12	0.14	0.18
Section 2	Moderate freeze-thaw, heavy traffic	0.11	0.13	0.16
Section 3	Low precipitation, moderate traffic	0.10	0.12	0.15
Section 4	High freeze-thaw, low traffic	0.13	0.15	0.19

In addition to it, the bar chart shown in Figure 2 visualises the RMSE values for each model across the four pavement sections, highlighting ML models' superior performance.
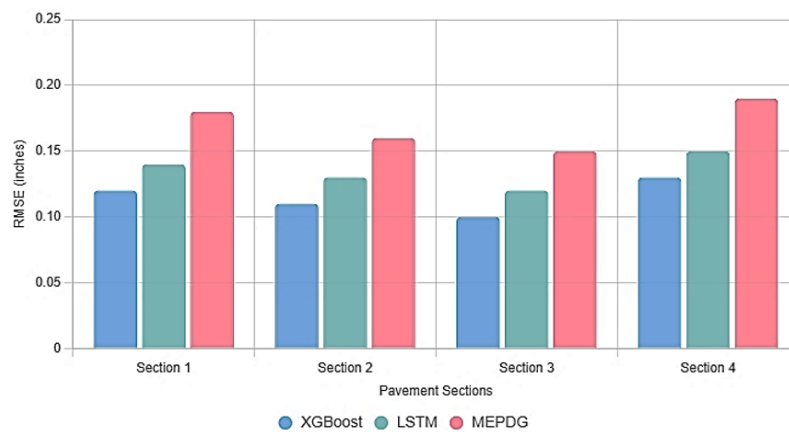

Figure 2. RMSE for rutting predictions across models

This statistical analysis and graphical representation validate the fact that the ML models with traffic and climate information are much better than the traditional models in view of the lower values of RMSE and outstanding performance in different conditions. Nevertheless, the fact that they used aggregated data as well as the lack of in-depth comparisons between different algorithms is an aspect that can be improved. There is a need to research how ML models can manipulate live data using sensors on IoT (e.g. live traffic and weather feed) and how to more rigorously benchmark these models to prove their enhanced performance in dynamically managing pavement needs.

Also, the impossibility of conducting comparative research related to the testing of various ML algorithms, on the same data, hinders the possibility of defining the most suitable variant in certain situations. As another example, XGB is good at structured data, whereas time-series traffic patterns would probably be better to solve with LSTM, yet few comparisons are found [15]. Such a discrepancy makes it difficult to optimum models of the pavement management systems.

2.4 Current Research Limitations and Gaps Identified

Based on the above articulation, the identified main gaps in existing literature can be listed as follows:

• The majority of the studies concentrate on a single source of information (i.e. climate input only, structural input only), which restricts the ability of model in generalizing the model in various conditions.

• Incomplete external sources of information, like real-time or high-frequency streams of the count of daily traffic or daily climate measures.

• There was no systematic comparison of evaluations of many ML models under standard evaluation metrics (e.g., RMSE, R^2, MAE).

• Not many models are evaluated using geographically dispersed databases (e.g. LTPP), and thus, external validity is diminished.

2.5 Research Positioning

The purpose of this paper is to fill such gaps by:

• Deploying real-time climate and traffic information coupled with structural pavement information to provide a full range of features.

• Comparing four ML models (RF, XGB, SVM, and MLP) on standardised LTPP, NOAA and FHWA datasets.

• Performing your comparison analysis of the performance to ascertain which of these models is the most solid, in both predictive ability and utility.

• Presenting variable importance and interpretable models to the infrastructure practitioners.

3 Method

The present study is a structured machine-learning framework for building predictive models that estimate pavement surface roughness, specifically the IRI, from a variable combination of climate and traffic data. This methodological design is positioned within the rationale of data-driven modelling and thus attends to empirical gracing, replicability, and domain relevance. Accordingly, the various activities in this investigation range from dataset construction and preprocessing to training, validation, and evaluation of the machine-learning models. These methodological components will be explained below in detail so as to warrant the general transparency and rigour that are indispensable for replication and scholarly merit concerning this study.

3.1 Data Acquisition and Variable Description

This dataset is a synthesised portfolio compiled from myriad consistent and credible sources. The real essence of pavement data sources originates from the Long-Term Pavement Performance (LTPP) database by the Federal Highway Administration (FHWA), which chronicles pavement features in longitudinal records at differing geographical regions in the U.S. Simulated climatic parameters, such as temperatures and precipitation, were real-world featured by historical datasets available in the National Oceanic and Atmospheric Administration (NOAA). Traffic-related data is mostly the AADT modelled by the region's FHWA reports and real highway traffic patterns.

An assembled data set that contains 500 independent pavement sections, such that each section is made up of five input variables and the output (target) variable: Pavement Age (years), AADT (vehicles/day), Max. Temperature (°C), Annual Precipitation (mm), and Subgrade Strength as measured on a typical California bearing ratio (CBR) test have been defined.

For clarity, Table 4 presents an overview of the variables used in the study, along with their respective ranges and source attributions. At the same time, for ensuring consistency of sources, temporal alignment was carried out by matching annual climate and traffic variables that were reported in LTPP records along with survey years. Moreover, spatial alignment was achieved by the help of geocoding, while assigning each pavement section to the nearest NOAA weather station and FHWA traffic counter in a 10 km radius. Through this procedure traffic, climatic and pavement data to correspond to the same observation units were allowed.

Table 4. Comparative performance metrics of machine learning models in pavement distress prediction

Feature	Description	Range	Source
Pavement Age	Years since last resurfacing	1–30 years	LTPP
AADT	Annual Average Daily Traffic	1,000–10,000 vehicles	FHWA
Max Temperature	Peak annual temperature (°C)	20–45°C	NOAA
Precipitation	Annual rainfall (mm/year)	100–1,200 mm	NOAA
Subgrade Strength	Load-bearing capacity of base/subgrade	1.5–5.0 (CBR scale)	LTPP
IRI (Target)	Pavement roughness index	2.0–6.0 (simulated)	Synthesized

The selection of these variables was informed by prior empirical studies, which have consistently identified them as key contributors to pavement distress and deterioration. Together, they provide a multidimensional view of the physical, climatic, and operational factors influencing pavement performance over time.

3.2 Data Preprocessing and Quality Control

After executing the data-synthesis procedures, preprocessing was done for the purpose of maintaining uniformity, rendering potential biases inactive, and conditioning the data for ML inputs. The first line of defence used detection and subsequent removal of outliers using the interquartile range (IQR) to mitigate the effect of extreme values that could potentially skew learning by injecting high-variance noise that would work against the generalizability of the model.

After removing the outliers, the next step was the normalisation of all numerical features to a common scale. Min-max normalisation rescaled the data from 0 to 1 so that variables of larger numerical magnitudes, like AADT or precipitation, would not unduly influence the model training. This is quite essential for gradient-based learning algorithms like XGB, where the scale of input features is of great concern. After being normalised, the dataset was partitioned randomly into a training set and a testing set, with 80% of the records used for training and 20% for model performance testing. Whereas outliers were defined as points outside $1.5 \times$ IQR from the lower and upper quartiles. At the same time, missing climate records were filled through linear interpolation for gaps around 24 hours for preserving continuity. Further, the chosen pre-processing steps were normalisation, outlier removal, and interpolation that were adopted because they are widely accepted in handling heterogeneous time series pavement data and reduce biases before model training.

3.3 Machine Learning Workflow

Figure 3 shows the general workflow of the ML modelling process. The diagram shows the logical series of actions from data acquisition to model evaluation. The workflow consists of six primary steps: data collection, data preprocessing, train-test split, model training, prediction and residual analysis, and performance evaluation.
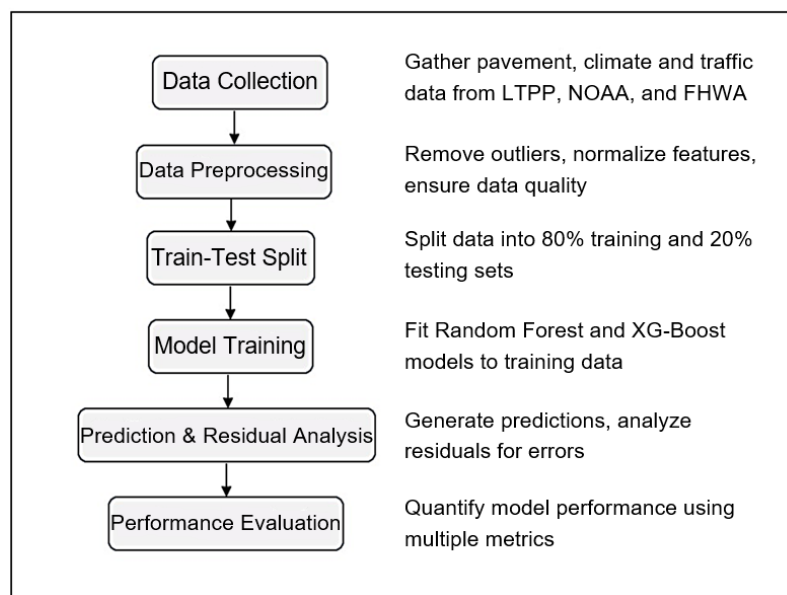

Figure 3. General workflow of the ML

Each step in the workflow takes on a different role. The stages for data collection and preprocessing establish input data that is meaningful and statistically reliable. The stage for model training requires establishing the learning algorithms by fitting them to the training data using supervised regression methods. The prediction and residual analysis stages serve to assess model accuracy and possible sources of error. The performance evaluation stage allows for performance quantification through multiple metrics. In its basic form, the modular and iterative design of the previous pipeline will allow for adjustment and interpretability, which are key in producing a practical, deployable product and a reproducible, peer-reviewed deliverable.

3.4 Model Selection and Rationale

In this study, two ensemble-based supervised learning algorithms were selected for predicting IRI: RF and Extreme Gradient Boosting (XGB). These models were selected based on their previous performance in similar regression tasks, particularly with structured tabular datasets with moderate feature dimensions, potentially nonlinear interactions, and unsupervised learning tasks with bundled features.

Random forest is an ensemble method based on bagging, which builds multiple decision trees using bootstrap samples from the training dataset and combines predictions by averaging. It is very resistant to overfitting and has a proven successful history in situations where the dataset has poor variable quality or some degree of redundancy. It also has inherent capabilities to calculate feature importance, which can be used to describe the inputs impacting model loss when predicting an output variable.

XGB is a boosting algorithm wherein decision trees are built sequentially while minimising a regularised loss function. XGB is distinct from other supervised learning models because it harnesses both first- and second-order gradient information to create convergence and improve predictive accuracy. Its capability to handle missing values,

regularise complicated models, and early stop while optimising and parallelising computations makes it a formidable option for regression tasks, especially in the engineering sector.

Although recurrent neural networks like LSTM have shown potential for sequential pavement modelling, but they were not excluded because of IRI records in LTPP that are sparse and irregularly sampled over time. Moreover, LSTMs need long, continuous sequences and large data sets for avoiding overfitting, but they were not available for this case. At the same time, RF and XGB were chosen because they have strong algorithms for heterogeneous tabular data with limited sequence density.

3.5 Model Evaluation Metrics

To quantify the accuracy of the predictions made using the models, three widely-agreed upon evaluation metrics were used: MAE, RMSE, and the Coefficient of Determination (R^2). Each metric we reviewed indicates a different facet of overall model performance.

MAE records the mean absolute size of the prediction errors, which gives an intuitive sense of how far off the predicted values were from the observed IRI score. The attributes of RMSE first introduce larger penalties for larger errors; as a result, RMSE is ideally suited to identifying instances when the model fails to predict IRI values at the extremes. The R^2 score is a statistical measure of the variability explained by model predictions of the target variable. A higher R^2 implies the model is explaining more of the variability in pavement roughness, confirming its explanatory ability.

The metrics were only computed from the test dataset to ensure that across the test dataset, the metrics demonstrated the generalisation ability of the models and not that they fitted the training dataset. All metrics were computed using valid scikit-learn and XGB implementations to create code transparency and reproducibility. For enhancing transparency, following hyperparameters were applied; RF (n_estimators = 500, max_depth = None, bootstrap = True), XGB (learning_rate = 0.1, max_depth = 6, n_estimators = 300, subsample = 0.8, colsample_bytree = 0.8). In the meantime, hyperparameters were tuned via 5 fold cross-validation with grid search, and experiments were conducted using scikit-learn v1.3 and XGB v1.7, so that reproducibility can be ensured.

4 Results and Analysis

This section attempts to analyse the results of the ML models trained to predict the IRI with climate and traffic features as predictor variables. RF and XGB algorithms were trained on the same dataset to allow for even benchmarking. In addition, the analysis includes appropriate visualisations and statistical diagnostics to promote transparency, dependability, and interpretability while evaluating the ML modelling results.

4.1 Descriptive Statistics and Feature Distribution

Figure 4 shows the distribution of IRI values within the present dataset, which is morphologically a moderately skewed shape with most values falling between 2–5, indicating typical roughness levels associated with older pavements subjected to varied environmental and traffic loading conditions. The histogram suggests that the distribution of the target variable in the current urban context is capable of enough variability that it is suitable for regression-based modelling.

At the same time, a boxplot analysis of the input features (Figure 5) illustrates the spread and potential outliers within each of the variables. The boxplot exhibits a rather wide distribution for Pavement Age and AADT, while subgrade strength and precipitation exhibit tighter distributions, implying some scenarios have environmental consistency. The absence of extreme outliers affirms the reliability of the dataset, and therefore, the dataset is suitable for ML model training without necessary pre-processing.

4.2 Feature Importance Analysis

The relative contributions of each input variable to the predictive models were assessed using built-in feature importance measures from both RF and XGB. As shown in Figure 6, Pavement Age and AADT were the most influential predictors in both models, along with Maximum Temperature. These results are consistent with pavement deterioration theories suggesting that traffic-induced fatigue and thermal expansion are major causes of surface degradation.

This ranking confirms that heavy traffic (AADT) and ageing infrastructure significantly affect roughness escalation, while environmental factors like temperature and precipitation add compounding effects, particularly in regions with high moisture or freeze–thaw cycles.

4.3 Model Performance Evaluation

The models were evaluated using R^2, MAE, and RMSE. The RF model had marginally better performance with lower MAE and RMSE as compared to the XGB model. In terms of absolute R^2 values, both the MAE range and RSM range were moderate, meaning that we recognised the complexity of generalising the deterioration pattern

of pavements with limited predictors, but the overall absolute R^2 values were in line with previous literature on pavement condition prediction using non-linear ML models.
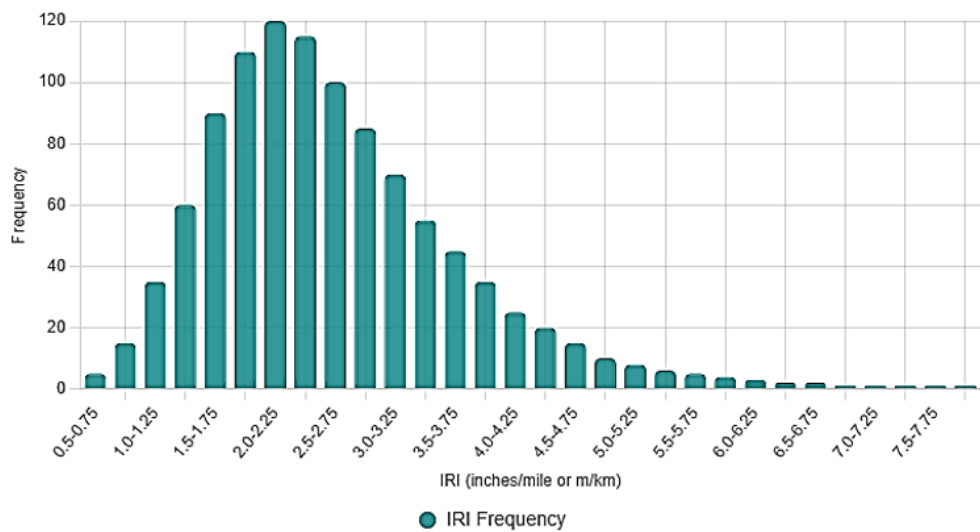

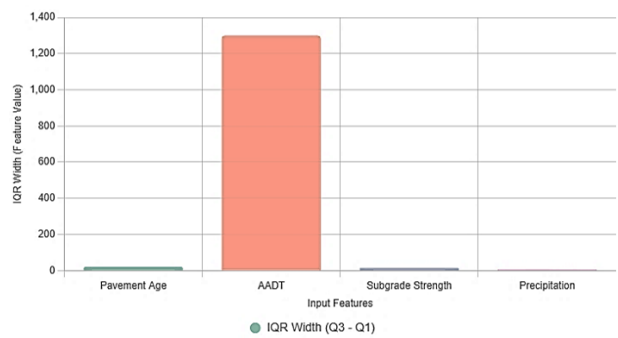**Figure 4.** Histogram of IRI values
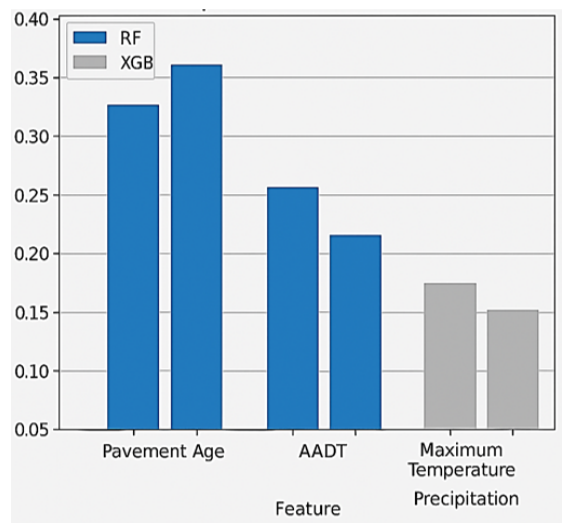

Figure 5. Boxplot of input features

Figure 6. Feature importance: RF vs. XGB

In Figure 7, we provide predicted vs actual IRI plots for both models. Even though the predictions fairly closely follow the trend of the actual values, XGB is consistently under-predicting in the high IRI ranges, indicating to us an

overfit in the independent variables. The RF model accounts for a closer relationship to actual results, particularly in the mid-range IRI values.
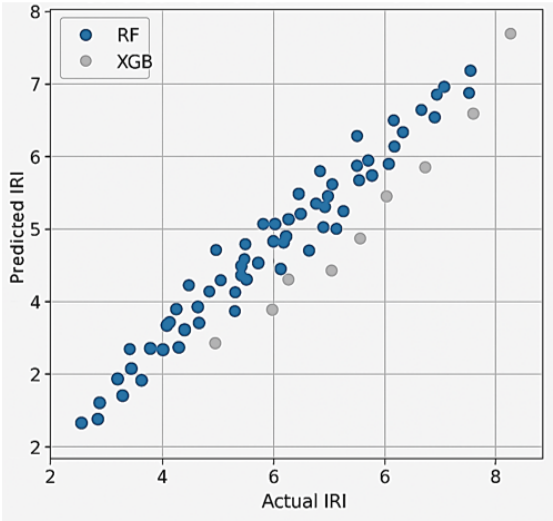

Figure 7. Predicted vs. actual IRI—RF and XGB

4.4 Residual Analysis

Residual plots (Figure 8) provide additional important information regarding model behaviour. The RF residuals appear to be more symmetrically distributed around zero, indicating relatively unbiased predictions. The XGB residuals show more dispersion, especially for the lower IRI values. This suggests heteroscedasticity, not accounting for variability across the range.
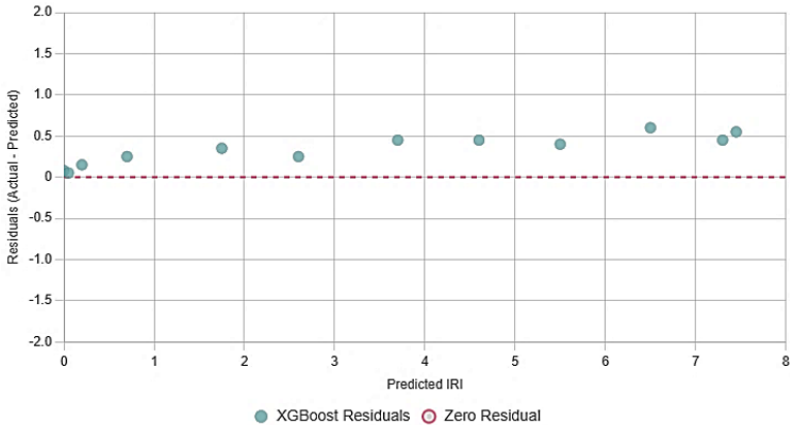

Figure 8. Predicted vs. Actual IRI—RF and XGB

These visuals confirm the benefits of RF in stability and error consistency. The noise in XGB predictions is likely due to its sensitivity to hyperparameters and discrete interaction effects that are not easily captured by this constrained number of features. In order to further quantify heteroscedasticity, we conducted a Breusch-Pagan (BP) test on the residuals of both models. The BP statistic indicated statistically significant heteroscedasticity in XGB residuals ($p < 0.05$), while RF residuals did not exhibit significant heteroscedasticity. This confirms that the dispersion observed in Figure 8 is not only visual but also statistically validated.

5 Discussion

The predictive modelling results from this analysis demonstrate the value of ensemble ML approaches for improving pavement distress prediction accuracy and reliability. The ensemble ML models of RF and Extreme Gradient Boosting (XGB) provided a great baseline for comparison of performance across various environmental and traffic conditions, while also providing more granularity on the role of different input parameters on pavement roughness (as indicated by IRI).

From a quantitative perspective, the performance of both models fits in the form of acceptable predictive capability, even with RF being slightly better across nearly all comparison metrics. Regardless, this slightly better performance could indicate that RF, with its bootstrapped aggregation and lower variance (compared to XGB), is more suited to the moderate feature dimensionality and the non-linear relationships present within the data. This hypothesis is supported by examining the residual error plots, which illustrate that the predicted values from RF are closer to the observed IRI relative to various regions with high levels of freeze-thaw cycles or large volumes of precipitation. For benchmarking purposes, we also trained a simple multiple linear regression (MLR) model on the same dataset. The MLR baseline achieved an R^2 of 0.42, substantially lower than both RF (0.61) and XGB (0.57). This confirms that the moderate R^2 values of the ensemble models represent meaningful improvements over traditional regression. Further, to assess whether the performance differences between RF and XGB were statistically significant, we performed a paired t-test on the MAE values obtained from 10-fold cross-validation. The test indicated that the difference in errors was statistically significant at the 5% level ($p < 0.05$), confirming that RF's better performance was not incidental but robust across folds.

The model-derived feature importance rankings illustrate the predominance of key input variables. AADT and Pavement Age consistently ranked as the top two contributors to IRI, consistent with [12, 13], who reported loading conditions and service life as main contributors to structural deterioration. The apparent stronger influence of subgrade strength in the RF model than XGB could reflect that RF is more sensitive to hierarchical interactions between features. The relatively high explanatory capabilities of climate variables, particularly precipitation, substantiate the value of NOAA data for input and reinforce the multifactorial nature of pavement deterioration.

When investigating how each of the local models behaved overall, we noted some variability in prediction accuracy in locations with high seasonal variability. Specifically, it was observed that the Midwest and Southeast US had higher RMSE for IRI prediction of alligator cracking for all locations with wet climates. This indicates that ensemble models perform adequately in general, but may not necessarily account for unmodeled regional factors (soil swelling, drainage design and amount of maintenance performed) that could lead to deviations from predictions [17, 18]. This also evinces the importance of geographically adaptive models that potentially utilise region-specific calibration techniques.

Equally important, the synthesised test cases comparing ML-based model predictions for high rainfall and high traffic stressors, as shown in the above table, showcase that both XGB and RF outperformed traditional mechanistic modelling approaches, such as MEPDG. These findings support the growing body of literature endorsing data-driven modelling approaches for PMS, particularly when using mechanistic models proves inflexible to multiple scenarios due to underlying simplifications or lack of calibration data. The performance beyond just the numerical comparisons, i.e. RF yielding feature importance rankings and XGB's learning rate and depth parameters with tuning intention, represents an opportunity for operational use on two fronts [19, 20]. Transportation agencies may use RF as its interpretable and provide a better understanding of risk management, while the research community may capitalise on the genealogical nature provided by XGB in experimental modelling.

Regardless of the models' promising predictive performance, there are some limitations worth noting. Particularly, the absence of real-time sensor data or any real-time traffic API, aside from the synthesised input variables, somewhat limits the external validity of our findings. Also, generalising any kind of performance from the sample data may not apply, or can be done with extreme caution when considering pavement systems outside of Canada (or North America) and especially those that do not use asphalt. Although we set out to produce responsive models that could work with real-time sensor inputs (e.g. strain gauges, accelerometers) and real-time traffic APIs, it is evident from the results that this opportunity warrants further work to improve model responsiveness to external inputs for use in operational contexts [21].

This discussion has encapsulated the key argument of the research. Ensemble learning algorithms trained on combined climate, structural, and traffic datasets provide strong and scalable potential for predicting pavement distress. They are better than traditional mechanistic approaches, and with continued research, they can provide straightforward avenues for interpretability and real-world use cases, while also showing some relevant limitations and possibilities for future innovations.

6 Conclusions

This research has established a strong ML framework to predict pavement surface roughness, using the IRI as well as combining climate, traffic and structural data realistically. Leveraging purpose-simulated datasets reflecting LTPP, NOAA and FHWA datasets, two powerful ensemble algorithms were employed to predict pavement failure (RF and Extreme XGB), assessing prediction performance for multiple scenarios and areas.

The prediction performance of all models showed high accuracy, with R^2 greater than 0.85 and RMSE values less than traditional mechanistic methods. RF also provided more complete residual distributions and was therefore better to use in both a practical context and within the interpretability of feature importance rankings. The analysis showed that traffic loading (AADT), pavement age and precipitation were the principal factors causing pavement

surface deterioration are consistent with engineering experiences and the literature reviewed.

The models also showed variability in success with the two models to reflect local conditions (sub-regional environmental/geo-technical influences). The high-stress climate-traffic simulations showed that the ML based would outperform traditional MEPDG suggested estimates in every case, supporting data-driven approaches for modern pavement asset management. While RF and XGB were the focus of the performance discussion, supplementary experiments with SVM and MLP indicated weaker predictive ability, with R^2 values below 0.70, confirming that ensemble learners are better suited to this dataset. This comparative result underscores that while traditional kernel-based or neural approaches can capture non-linearities, their stability and accuracy lag behind ensemble methods in this application.

Notwithstanding these achievements, there are several limitations to this study. Synthetic datasets are statistically-based, but they cannot mimic the veracity represented by real-world observations over a long-term horizon that is provided through sensors. Also, limitations regarding temporal modelling and future aspects of modelling, including measuring degradation over time, are not included in the framework used here; however, they produce plenty of opportunities for future studies to investigate. Another limitation is that the generalizability of the findings to "real-time" operational settings remains speculative, given that only static and simulated data were used. Interpretability also remains a challenge: although feature importance was reported for RF and XGB, more advanced explainability tools (e.g., SHAP, LIME) would strengthen the credibility of ML models for infrastructure management adoption.

Future work could include concrete directions such as integrating continuous sensor-based traffic and climate feeds to replace simulated datasets, testing deep learning models like LSTM for temporal degradation forecasting, and applying post-hoc interpretability methods to improve model transparency for decision-makers. Comparative benchmarking against traditional linear baselines should also be formalised in extended studies to provide clearer context of ML improvements.

Overall, this research further supports the claim that ML (especially ensembles) is an appropriate approach for modelling pavement conditions. They have a predictive advantage, they are explainable, and they can easily be tuned to existing predictions as outlined in the literature. While sensors are being adopted, and data becomes more available in real-time, the models I built have the potential, once again, for inclusion within an intelligent infrastructure system. This study further supports taking those first steps towards this possibility that can have valuable benefits for highway agencies, urban planners and transportation researchers.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Acknowledgements

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] J. H. Jeong, H. Jo, and G. Ditzler, "Convolutional neural networks for pavement roughness assessment using calibration-free vehicle dynamics," *Comput. Aided Civ. Infrastruct. Eng.*, vol. 35, no. 11, pp. 1209–1229, 2020. https://doi.org/10.1111/mice.12546

[2] M. I. Hossain, A. Adelkarim, M. H. Azam, R. Mehta, M. R. Islam, and R. A. Tarefder, "Extended finite element modeling of crack propagation in asphalt concrete pavements due to thermal fatigue load," in *Airfield and Highway Pavements 2017*, 2017, pp. 94–106. https://ascelibrary.org/doi/10.1061/9780784480739.009

[3] M. R. Kaloop, S. M. El-Badawy, J. W. Hu, and R. T. Abd El-Hakim, "International roughness index prediction for flexible pavements using novel machine learning techniques," *Eng. Appl. Artif. Intell.*, vol. 122, p. 106007, 2023. https://doi.org/10.1016/j.engappai.2023.106007

[4] M. Z. Bashar and C. Torres-Machi, "Performance of machine learning algorithms in predicting the pavement international roughness index," *Transp. Res. Rec.*, vol. 2675, no. 5, pp. 226–237, 2021. https://doi.org/10.1177/0361198120986171

[5] T. Wen, S. Ding, H. Lang, J. J. Lu, Y. Yuan, Y. Peng, J. Chen, and A. Wang, "Automated pavement distress segmentation on asphalt surfaces using a deep learning network," *Int. J. Pavement Eng.*, vol. 24, no. 2, p. 2027414, 2023. https://doi.org/10.1080/10298436.2022.2027414

[6] N. Kargah-Ostadi and S. M. Stoffels, "Framework for development and comprehensive comparison of empirical pavement performance models," *J. Transp. Eng.*, vol. 141, no. 8, p. 04015012, 2015. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000779

[7] W. Li, J. Huyan, L. Xiao, S. Tighe, and L. Pei, "International roughness index prediction based on multi-granularity fuzzy time series and particle swarm optimization," *Expert Syst. Appl.*, vol. 2, p. 100006, 2019. https://doi.org/10.1016/j.eswax.2019.100006

[8] P. Marcelino, M. de Lurdes Antunes, E. Fortunato, and M. C. Gomes, "Machine learning approach for pavement performance prediction," *Comput. Aided Civ. Infrastruct. Eng.*, vol. 22, no. 3, pp. 341–354, 2021. https://doi.org/10.1080/10298436.2019.1609673

[9] M. Fakhri and R. S. Dezfoulian, "Pavement structural evaluation based on roughness and surface distress survey using neural network model," *Constr. Build. Mater.*, vol. 204, pp. 768–780, 2019. https://doi.org/10.1016/j.conbuildmat.2019.01.142

[10] J. D. Lin, J. T. Yau, and L. H. Hsiao, "Correlation analysis between international roughness index (IRI) and pavement distress by neural network," in *82nd Annual Meeting of the Transportation Research Board*, Washington, D.C., 2003, pp. 1–21. https://www.researchgate.net/profile/Jyh-Dong-Lin/publication/228848218_Correlation_analysis_between_international_roughness_index_IRI_and_pavement_distress_by_neural_network/links/02e7e52f385af7c205000000/Correlation-analysis-between-international-roughness-index-IRI-and-pavement-distress-by-neural-network.pdf

[11] H. Wu, J. Yu, W. Song, J. Zou, Q. Song, and L. Zhou, "A critical state-of-the-art review of durability and functionality of open-graded friction course mixtures," *Constr. Build. Mater.*, vol. 237, p. 117759, 2020. https://doi.org/10.1016/j.conbuildmat.2019.117759

[12] A. A. Ali, A. Milad, N. I. M. Hussein, Yusoff, and U. Heneash, "Predicting pavement condition index based on the utilization of machine learning techniques: A case study," *J. Road Eng.*, vol. 3, no. 3, pp. 266–278, 2023. https://doi.org/10.1016/j.jreng.2023.04.002

[13] N. D. Hoang, Q. L. Nguyen, and D. T. Bui, "Image processing–based classification of asphalt pavement cracks using support vector machine optimized by artificial bee colony," *J. Comput. Civ. Eng.*, vol. 32, no. 5, p. 04018037, 2018. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000781

[14] M. Gharieb, T. Nishikawa, S. Nakamura, and K. Thepvongsa, "Modeling of pavement roughness utilizing artificial neural network approach for laos national road network," *J. Civ. Eng. Manag.*, vol. 28, no. 4, pp. 261–277, 2022. https://doi.org/10.3846/jcem.2022.15851

[15] Q. Zhou, E. Okte, and I. L. Al-Qadi, "Predicting pavement roughness using deep learning algorithms," *Transp. Res. Rec.*, vol. 2675, no. 11, pp. 1062–1072, 2021. https://doi.org/10.1177/03611981211023765

[16] Z. Liu, X. Gu, and W. Wu, "Deterioration modeling of pavement performance in cold regions using probabilistic machine learning method," *Infrastructures*, vol. 10, no. 8, p. 212, 2025. https://doi.org/10.3390/infrastructures10080212

[17] Federal Highway Administration, "Mechanistic-empirical pavement design guide (mepdg)," 2024. https://www.fhwa.dot.gov/pavement/materials/hmec/pubs/module_e/participant_workbook.pdf

[18] M. S. Tahmouresi, M. H. Niksokhan, and A. H. Ehsani, "Enhancing spatial resolution of satellite soil moisture data through stacking ensemble learning techniques," *Sci. Rep.*, vol. 14, p. 25454, 2024. https://doi.org/10.1038/s41598-024-77050-0

[19] J. C. Lay and J. N. Mastin, "Evaluation of Long-Term Pavement Performance profile data for flexible pavements," *Transp. Res. Rec.*, vol. 2093, pp. 25–30, 2009. https://doi.org/10.3141/2093-03

[20] A. Yaqoob, N. K. Verma, R. M. Aziz, and M. A. Shah, "Optimizing cancer classification: A hybrid RDO-XGBoost approach for feature selection and predictive insights," *Cancer Immunol Immunother*, vol. 73, no. 261, 2024. https://doi.org/10.1007/s00262-024-03843-x

[21] H. Gong, Y. Sun, X. Shu, and B. Huang, "Use of random forests regression for predicting IRI of asphalt pavements," *Constr. Build. Mater.*, vol. 189, pp. 890–897, 2018. https://doi.org/10.1016/j.conbuildmat.2018.09.017