IRI Prediction using Machine Learning Models

ANKIT SHARMA, PRAVEEN AGGARWAL
Department of Civil Engineering,
National Institute of Technology Kurukshetra, Haryana-136119,
INDIA

Abstract: - Road infrastructure is the backbone of the economy of any country. The recent increase in the length of roads has never been matched in history. The increase in length comes with huge demand for the maintenance of pavements in an orderly fashion. The pavement management system is used for planning maintenance based on pavement performance evaluation. The international roughness index (IRI) is considered a standard parameter for the functional evaluation of flexible pavements. In the present study, IRI is predicted through machine learning models using the LTPP database. The main objective of the study is to find the optimal machine learning which can be used for IRI prediction. Three machine learning models, (i) linear regression, (ii) optimised trees, and (iii) optimised Gaussian process regression (GPR), has been used for predicting IRI. Different models have been compared based on various statistical parameters. The optimised GPR model performed best per the R-Squared value (0.89).

Key-Words: - International roughness Index, Pavement performance evaluation, Machine learning

Received: June 22, 2022. Revised: May 12, 2023. Accepted: June 15, 2023. Published: July 10, 2023.

1 Introduction

India is a vast country with a huge road network with a total road length of 63.86 lakh kilometers in 2019. The road length has increased at a compound annual growth rate of 4.2% since 1951. Rural roads constitute a major portion of the road network in India (approximately 71%). The huge network of rural roads needs regular maintenance. Hence pavement evaluation is vital for decision-makers to preserve the road infrastructure and to maximise its benefits to the public.

Researchers have conducted multiple studies to conclude the importance of distress in quantifying pavement quality, as the road length is increasing per year at a fast pace. The already available infrastructure needs to be simultaneously evaluated for maintenance requirements. The limited budgets and increasing public expectations stressed the decision-makers in adopting a temporary solution to pavement distress. The inter-correlation between the pavement distresses is often ignored. The correlation analysis of multiple pavement sections will help understand the requirement for treating underlying problems. The decision-makers can then provide adequate treatment accordingly. An ideal Pavement management system (PMS) should hence integrate the evaluation of all distresses simultaneously.

Visual inspection of various pavement distresses is adopted by in-field engineers for assessment of the current condition of the pavement. This ought to be a very initial part of the pavement evaluation and needs quantitative measurements of ride quality/ pavement unevenness for judging the maintenance requirement of the road. The international roughness index (IRI) is used in evaluating the ride quality of the road. The IRI has been widely adopted by transportation agencies worldwide as a key indicator for evaluating pavement ride quality. However, the measurement of IRI is costly and unaffordable for highway agencies limited by their fiscal resources.

The recent advent of machine learning methods has enabled researchers to explore the applicability of machine learning field to their respective fields. Many past studies have been performed for the evaluation of the applicability of machine learning in the transportation engineering field, [1], [2], [3], [4].

In this study, data has been collected for 211 pavements from the LTPP database, and machine learning methods, i.e., linear regression, Optimised trees, and Optimised gaussian process regressions (GPR), have been explored. The results have been compared based on different model performance matrices, i.e., Root Mean Square Error (RMSE), R-squared, Mean Square Error (MSE), and Mean Average error (MAE). MATLAB computer program, [5], has been used in this study to achieve the desired results. The optimised GPR performed

the best in terms of all the model's performance values.

The manuscript is arranged into 5 Chapters in the order: Chapter 1: Introduction, Chapter 2: Data Preparation, Chapter 3: Machine Learning Models, Chapter 4: Methodology, Chapter 5: Results, and Chapter 6: Conclusions.

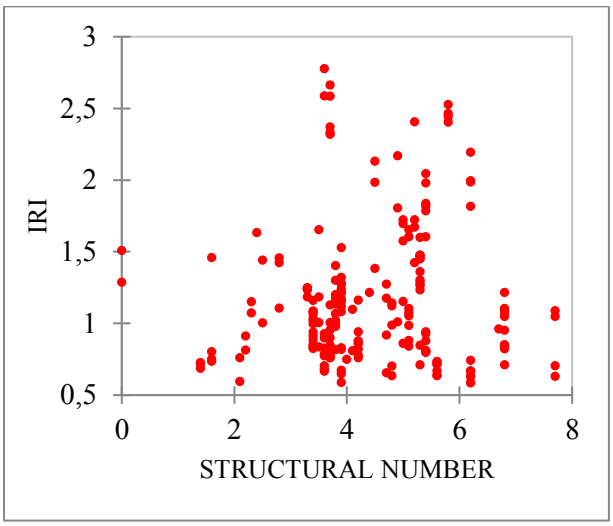
2 Data Preparation

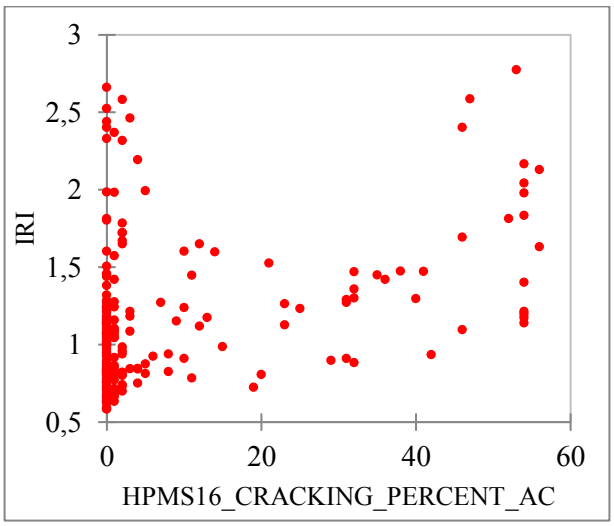
The data is extracted from the LTPP infopave database provided by Federal Highway Administration (FHWA). The database is accessible on the internet [6]. Table 1 describes the various variables retrieved for conducting the study. The data were extracted for all the pavement sections with flexible pavement surfaces.

Table 1. Name and description of different variables under consideration

NAME OF VARIABLE	DESCRIPTION OF VARIABLE
STRUCTURAL NUMBER	The structural number of asphalt concrete pavements.
HPMS16_CRACKING_PERCENT_AC	Total percentage of sections cracked by HPMS field guide definitions.
MEPDG_CRACKING_PERCENT_AC	This variable represents the sum of alligator cracks in percentage, per MEPDG, [7].
MEPDG_CRACKING_LENGTH_AC	This variable defines transverse cracks per meter run of the road via MEPDG definitions.
MEPDG_LONG_CRACK_LENGTH_AC	The length of longitudinal cracks conforms to the wheel path per meter length.
ME_PERCENT_WHEEL_PATH_CRACK	Percent of area cracked in 0.61-meter wide wheel paths as per MEPDG crack percent.
kesal_year	It defines the computed ESAL (Equivalent Single Wheel Load) for the k^{th} year.
unbound granular base thickness average	The average thickness of the unbound granular base layer.
asphalt concrete thickness average	The average thickness of the asphalt concrete layer.
MEAN_ANN_TEMP_AVG	It defines the average mean temperature for the whole year.
TOTAL_ANN_PRECIP	The sum of precipitation values for the whole year.
initial IRI for pavement	IRI value for the pavement in the year of construction.
AGE	This variable defines the age of the pavement in years.
IRI	International Roughness Index

The data extracted has been pre-processed using Microsoft Excel. The outliers were identified and removed to maintain the uniformity of the data. The scatter plots have been prepared for all the independent variables with dependent variables (Figure 1).





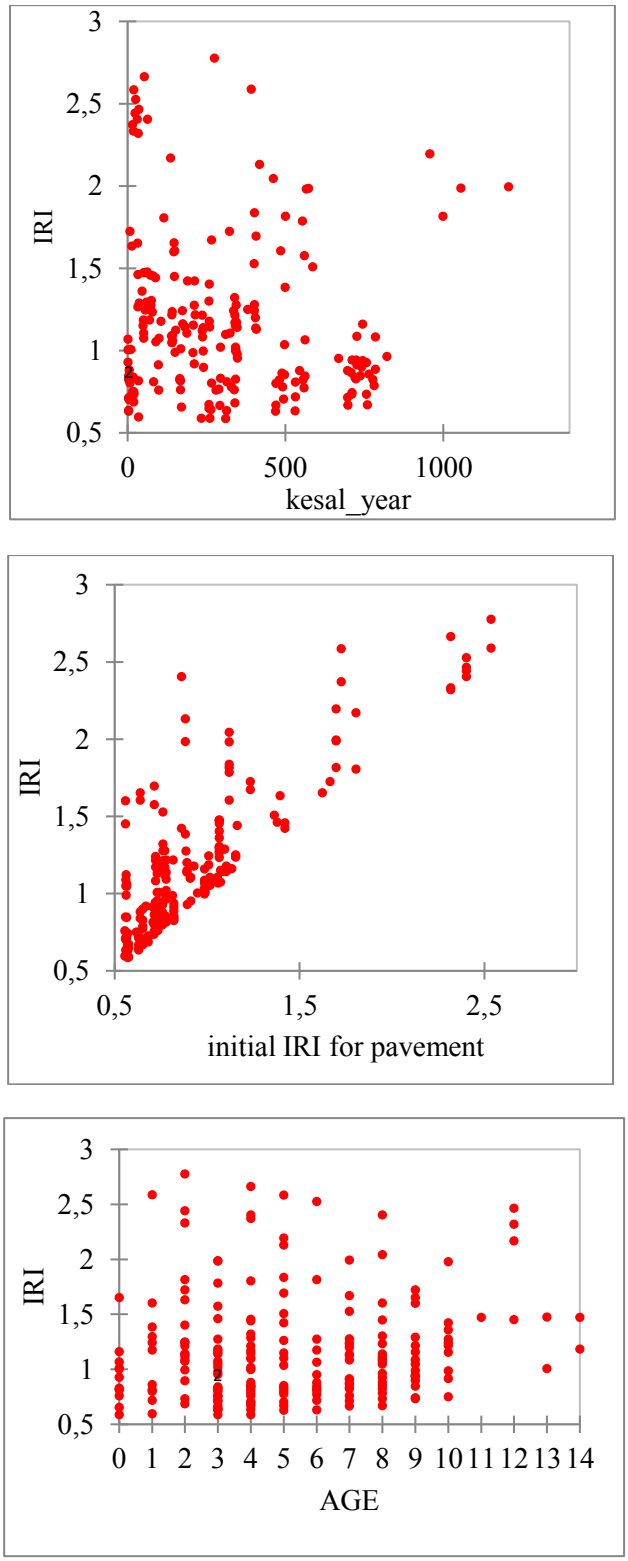


Fig. 1: Variation of IRI with different parameters under consideration

3 Machine Learning Models

In the present study, three machine models avail in MATLAB are used.

3.1 Linear Regression

Linear regression models describe the relationship between the dependent variable y and independent variables X. The matrix X is usually named as "Design matrix" of independent variables. The dependent variable is also named the response variable. A general equation for any line regression model has been given below (equation 1).

$$y_i=\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_p X_{ip} \quad (1)$$
$$, i = 1 , \dots , n$$

where

- n, represent the number of observations
- y_i, represents the i^{th} response
- β_0 represents the constant term in the linear regression model
- $\beta_1, \beta_2, \dots \beta_n$ represents the coefficients for all the independent variables
- $X_{i1}, X_{i2}, \dots , X_{ip}$ represent different independent variables for the ith response.

The method of least squares is used in this study to develop regression models. The method of least squares is used to determine the best-fit line for the independent variables.

3.2 Optimised Trees

In this study, optimised trees were used for the modelling. Decision trees are generally used for classification tasks in the machine-learning field. However, for regression analysis of the problem at hand, they can be used by converting the available data into small classification tasks. Figure 2 represents a typical decision tree. In this study, the decision trees have been optimised using Bayesian optimisation and hence are here referred to as optimised trees.

The decision of the models is based on the rules as defined in Figure 2. The figure describes the route to follow to determine if a person is fit. In the first step arithmetic operation on the age of the input variable age determines which route the model has to follow. Step 2, based on the input response model, determines if the input variable to be compared is "Eats a lot of pizza" or "Exercises in the morning". The answer to the questions decides if the person is fit or unfit.

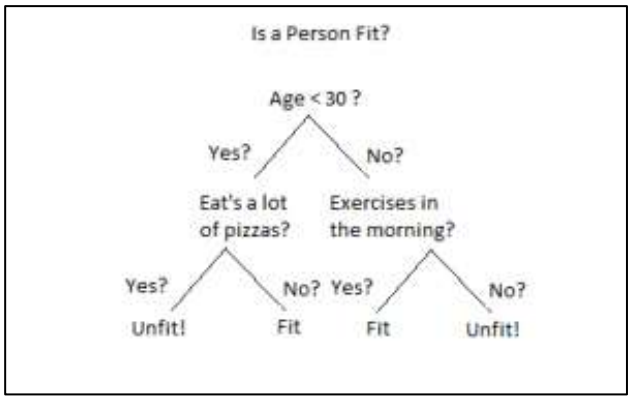


Fig. 2: Example decision tree

3.3 Gaussian Process Regression

It is a non-parametric Bayesian approach to solving regression problems. It is able to widely capture the relationship between predictors and responses. The complexity of the data is mapped using the kernel functions. The kernel function quantifies the similarity between the input variables. The similarity between the variables is used to generate the covariance matrix. The covariance defines how the output of the model is correlated to the input variables. Bayesian optimisation has been used for choosing the best hyperparameters for which the model performs best.

3.4 Optimisation

Optimisation of a machine learning model has to be performed when the parameters defined for the model have large sets, and many possible combinations of the hyperparameters are possible. The Bayesian optimisation technique is used to look for the best possible values of the input hyperparameters for the machine learning model. This optimisation is preferred when the selection of hyperparameters is huge, and it is not computationally possible to find the best-performing hyperparameters in the polynomial time. One of the key benefits of Bayesian optimisation ought to be its capability to balance the trade-off between exploration and exploitation in the available solution space. It uses a probabilistic model of the function to guide the search for optimal solution and balance the exploration of new inputs and exploitation of known good inputs.

4 Methodology

The application of machine learning algorithms has been done in the MATLAB program. The descriptive statistics of the variables were analysed to find out the different machine learning algorithms that can be applied to predict IRI accurately. The Regression learner has been used for the application of linear regression, optimised trees, and optimised GPR. To validate the results, the Cross-validation method of validation has been used along with 5-fold cross-validation. The data was divided into five equal but random parts by the machine learning models and was used for improving the training of the model.

The Bayesian optimisation technique was selected for the optimisation of the hyperparameters for different models. The trained models were then compared based on the model performance indices.

5 Results

The performance of all the models has been compared to each other in terms of performance parameters as discussed in, [8], [9], [10], [11]:

- Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (2)$$

- Mean Average Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

- R-squared:

$$R^2 = 1 - \frac{\sum_{i=1}(\hat{y}_i - y_i)^2}{\sum_{i=1}(\hat{y}_i - \bar{y})^2} \quad (4)$$

- Mean Square Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (5)$$

$\hat{y}_i$ = predicted values by the model
y_i= observed values for those inputs
n= total number of observations
$\bar{y}$= average of the measured values.

Table 2. depicts the parameters for the developed models

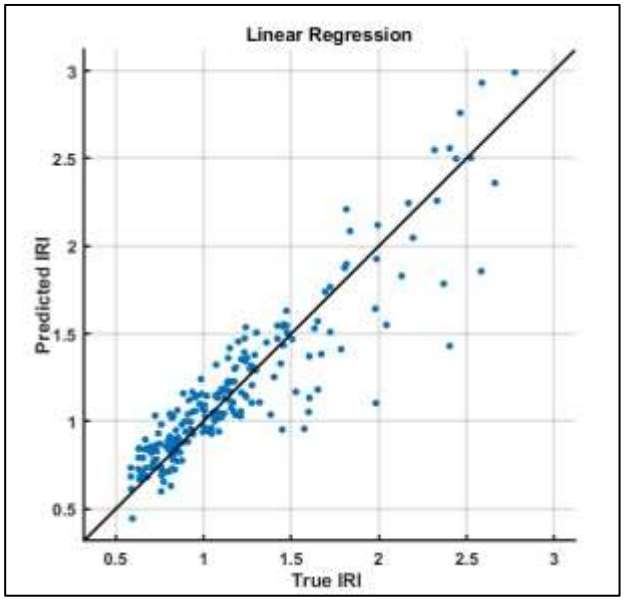
	RMSE	*R-squared*	*MSE*	*MAE*
Linear regression	0.1976	0.83	0.039	0.13
Optimised Trees	0.2182	0.80	0.047	0.14
Gaussian Process Regression	0.1576	0.89	0.024	0.09

From Table 2, it can be observed that the optimised trees have performed the lowest of all the models trained. The Classification and regression trees algorithm, [12], performs better when the data is not complex, and the data to be modelled is categorical variables. In the dataset of pavement, the dataset is in tabular format, and optimised trees are hence performing badly in performance parameters.

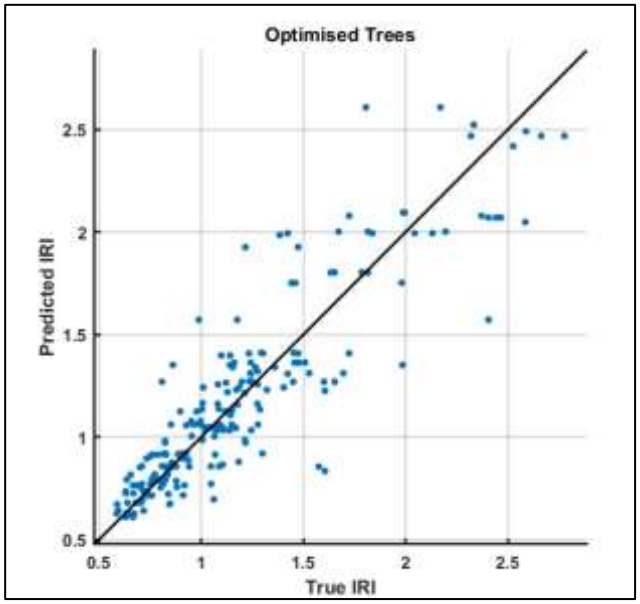
The scatter plot for actual and predicted values from the validation results from linear regression has been displayed in Figure 3 (a).

The linear regression model performed better than optimised trees with an r-squared value (0.83). The root mean square value for the model was 0.1976. actual vs predicted scatter plot for the validation dataset has been shown in Figure 3(a).

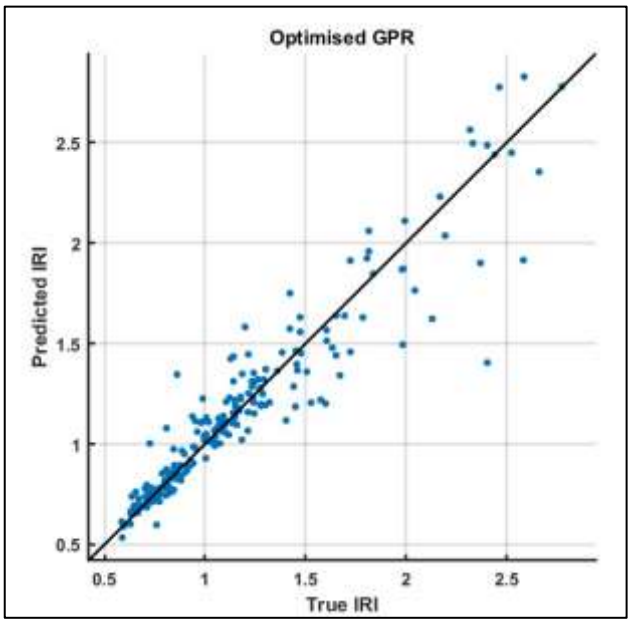
The optimised GPR model performed best with an R-squared value of 0.89. The performance of the model was at par with the previous studies conducted in the prediction of [3], [13], [14], [15]. Figure 3(c) represents the scatter plot between actual and predicted IRI values from the optimised GPR model.



(a) Actual vs predicted linear regression



(b) Actual vs predicted Optimised trees



(c) Actual vs predicted for Optimised GPR

Fig. 3: True versus predicted IRI value

6 Conclusion

This study analyses the applicability of machine learning models to the prediction of the road roughness of the pavement using various input parameters, i.e. structural number, HPMS16_CRACKING_PERCENT_AC, MEPDG_CRACKING_PERCENT_AC, MEPDG_CRACKING_LENGTH_AC, MEPDG_LONG_CRACK_LENGTH_AC, ME_PERCENT_WHEEL_PATH_CRACK, kesal_year, unbound granular base thickness average, asphalt concrete thickness average, MEAN_ANN_TEMP_AVG, TOTAL_ANN_PRECIP, initial IRI for pavement, and AGE. The different models that were utilised in this study are linear regression, optimised trees, and optimised GPR. The different performance indices used are RMSE, R-square, MAE, and MSE. The following conclusions are drawn from the study and listed below:

1. The use of the Bayesian method yields machine learning models that perform at par with past studies. The use of Bayesian optimisation enables the user to automatically search for optimised hyperparameters for the model based on the performance indices.
2. In terms of the R-squared value of the predictions, the model's optimised trees performed the lowest, and optimised GPR performed the best.

3. From the scatter plot of actual vs predicted, it is evident that the predictions by optimised GPR are close to the actual observed values for IRI. Hence optimised GPR model can be effectively used for accurate prediction of IRI using given parameters.

References:

[1] Inkoom S, Sobanjo J, Barbu A, Niu X (2019) Pavement Crack Rating Using Machine Learning Frameworks: Partitioning, Bootstrap Forest, Boosted Trees, Naïve Bayes, and K - Nearest Neighbors. *J Transp Eng Part B Pavements* 145:04019031. https://doi.org/10.1061/JPEODX.0000126.

[2] Marcelino P, de Lurdes Antunes M, Fortunato E, Gomes MC (2020) Transfer learning for pavement performance prediction. *Int J Pavement Res Technol* 13:154–167. https://doi.org/10.1007/s42947-019-0096-z.

[3] Zeiada W, Dabous SA, Hamad K, et al (2020) Machine Learning for Pavement Performance Modelling in Warm Climate Regions. *Arab J Sci Eng* 45:4091–4109. https://doi.org/10.1007/s13369-020-04398-6.

[4] Karballaeezadeh N, Danial MS, Moazemi D, et al (2020) Smart structural health monitoring of flexible pavements using machine learning methods. Coatings 10:1–18. https://doi.org/doi.org/10.20944/preprints202004.0029.v1.

[5] The MathWorks Inc. (2022). MATLAB version: 9.12.0.2039608 (R2022a) Update 5, Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com (Accessed: March 25, 2019).

[6] E. Elkins G, Ostrom B (2019) Long-Term Pavement Performance Information Management System User Guide, LTPP_IMS_USER_GUIDE_2019_V8.pdf , Available at : https://infopave.fhwa.dot.gov/InfoPave_Repository/files/LTPP_IMS_USER_GUIDE_2019_V8.pdf.

[7] Transportation Officials. (2008). Mechanistic-empirical pavement design guide: a manual of practice. AASHTO, available at : https://fenix.tecnico.ulisboa.pt/downloadFile/563568428712666/AASHTO08.pdf (Accessed: March 30, 2019).

[8] Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. Geosci Model Dev. https://doi.org/10.5194/gmd-7-1247-2014.

[9] Botchkarev A (2019) A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. *Interdiscipl. J Information, Knowledge, Manag* 14:045–076. https://doi.org/10.28945/4184.

[10] Krijnen WP (2006) Some results on mean square error for factor score prediction. Psychometrika. https://doi.org/10.1007/s11336-004-1220-7.

[11] Miles J (2014) *R Squared, Adjusted R Squared*. In: Wiley StatsRef: Statistics Reference Online

[12] Ziegler A, König IR (2014) Mining data with random forests: *Current options for real-world applications*. Wiley Interdiscip Rev Data Min Knowl Discov. https://doi.org/10.1002/widm.1114.

[13] Sharma A, Sachdeva SN, Aggarwal P (2021) Predicting IRI Using Machine Learning Techniques. *Int J Pavement Res Technol*. https://doi.org/10.1007/s42947-021-00119-w

[14] Sollazzo G, Fwa TF, Bosurgi G (2017) An ANN model to correlate roughness and structural performance in asphalt pavements. *Constr Build Mater* 134:684–693. https://doi.org/10.1016/j.conbuildmat.2016.12.186.

[15] Abdelaziz N, Abd El-Hakim RT, El-Badawy SM, Afify HA (2020) International Roughness Index prediction model for flexible pavements. *Int J Pavement Eng* 21:88–99. https://doi.org/10.1080/10298436.2018.1441414.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

-Ankit Sharma collected data and carried out model preparation

-Praveen Aggarwal helped in finalising the manuscript

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No external source of funding

Conflict of Interest

The authors have no conflict of interest to declare.