Transportation Consortium of South-Central States

Tran-SET

Solving Emerging Transportation Resiliency, Sustainability, and Economic Challenges through the Use of Innovative Materials and Construction Methods: From Research to Implementation

Development of Distress Index Prediction Models for Rehabilitation Treatments in Louisiana Using Advanced Machine Learning Techniques

Project No. 21PLSU15

Lead University: Louisiana State University

Final Report

August 2022

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation’s University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

Acknowledgements

The authors would like to acknowledge the support by the Transportation Consortium of South-Central States (Tran-SET) for supporting us throughout the project.

TECHNICAL DOCUMENTATION PAGE

1. Project No. 21PLSU15	**2. Government Accession No.**	**3. Recipient's Catalog No.**

4. Title and Subtitle Development of Distress Index Prediction Models for Rehabilitation Treatments in Louisiana Using Advanced Machine Learning Techniques	**5. Report Date** Aug. 2022 **6. Performing Organization Code**
7. Author(s) PI: Momen R. Mousa https://orcid.org/0000-0002-1723-364X Co-PI: Marwa Hassan https://orcid.org/0000-0001-8087-8232	**8. Performing Organization Report No.**
9. Performing Organization Name and Address Transportation Consortium of South-Central States (Tran-SET) University Transportation Center for Region 6 3319 Patrick F. Taylor Hall, Louisiana State University, Baton Rouge, LA 70803	**10. Work Unit No. (TRAIS)** **11. Contract or Grant No.** 69A3551747106
12. Sponsoring Agency Name and Address United States of America Department of Transportation Research and Innovative Technology Administration	**13. Type of Report and Period Covered** Final Research Report Aug. 2021 – Aug. 2022 **14. Sponsoring Agency Code**

15. Supplementary Notes

Report uploaded and accessible at Tran-SET's website (http://transet.lsu.edu/).

16. Abstract

Performance prediction models are used by state agencies to predict future trends in distress indices, hence, determining the required maintenance and/or rehabilitation treatment as well as the deterioration rate and remaining pavement service life. However, most of these models are based on a limited number of parameters and cannot predict the performance distress indices reliably. Such limitation resulted in having, most of the time, a maximum prediction period of five years. As a solution and coping with the ever-increasing size of pavement data, machine learning techniques have become a promising alternative. The objective of this study was to develop a machine-learning-based framework for states with a hot and humid climate that can predict the long-term field performance (for 11 years) of their asphalt (AC) overlays based on their key project conditions. Two machine learning algorithms were examined, namely Random Forest (RF) and CatBoost, and the one yielding a higher accuracy was considered. In this study, the well-known pavement condition index (PCI) was used as the pavement performance indicator. A total of 892 log miles of AC overlay data were obtained from the Louisiana Department of Transportation and Development (LaDOTD) Pavement Management System (PMS) database. Based on the collected data, six models were trained (for each algorithm) and validated to predict the future PCI of AC overlays for up to 11 years. Results indicated that the RF algorithm yielded higher accuracy than the CatBoost Algorithm and thus the RF-based models were considered in the proposed decision-making framework.

17. Key Words Pavement Performance, Pavement Condition Index, Machine Learning Algorithms, CatBoost.	**18. Distribution Statement** No restrictions. This document is available through the National Technical Information Service, Springfield, VA 22161.

19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 33	**22. Price**

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized.**

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
		NOTE: volumes greater than 1000 L shall be shown in m^3		
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
^{o}F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	^{o}C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
^{o}C	Celsius	1.8C+32	Fahrenheit	^{o}F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS, ABBREVIATIONS, AND SYMBOLS

AC	Asphalt
LaDOTD	Louisiana Department of Transportation and Development
DOT	Department of Transportation
PCI	Pavement Condition Index
RCI	Random Cracking Index
PMS	Pavement Management System

EXECUTIVE SUMMARY

Performance prediction models are used by state agencies to predict future trends in distress indices, hence, determining the required maintenance and/or rehabilitation treatment as well as the deterioration rate and remaining pavement service life. However, most of these models are based on a limited number of parameters and cannot predict the performance distress indices reliably. Such limitation resulted in having, most of the time, a maximum prediction period of five years. As a solution and coping with the ever-increasing size of pavement data, machine learning techniques have become a promising alternative. The objective of this study was to develop a machine-learning-based framework for states with a hot and humid climate that can predict the long-term field performance (for 11 years) of their asphalt (AC) overlays based on their key project conditions. Two machine learning algorithms were examined, namely Random Forest (RF) and CatBoost, and the one yielding a higher accuracy was considered. In this study, the well-known pavement condition index (PCI) was used as the pavement performance indicator. A total of 892 log miles of AC overlay data were obtained from the Louisiana Department of Transportation and Development (LaDOTD) Pavement Management System (PMS) database. Based on the collected data, six models were trained (for each algorithm) and validated to predict the future PCI of AC overlays for up to 11 years. Results indicated that the RF algorithm yielded higher accuracy than the CatBoost Algorithm and thus the RF-based models were considered in the proposed decision-making framework.

1. INTRODUCTION

Asphalt (AC) overlays are generally used to restore the structural capacity of aged pavements due to their effectiveness in improving pavement condition while minimizing user delay *(1*). By assessing the present condition and future performance of asphalt overlays, several agencies guarantee significant budget savings (*2*). The performance of AC overlays is identified by measuring and observing their condition over their lifetime (*3*). Many indices such as the Pavement Condition Index (PCI), International Roughness Index (IRI), Random Cracking Index (RCI) are used to characterize the pavement condition. A comprehensive review of traditional pavement index prediction models showed that they are mechanistic-empirical or purely empirical in nature (*4*). In Louisiana, the distress index prediction models are mostly polynomial, power, exponential, or linear transformation functions of pavement age (*5*). These index prediction models have a certain basic statistical structure, specific assumptions, and certain relationships between the input and output variables which violates some imperative assumptions such as independence of the input variables for parametric methods. Such challenge weakens the statistical power of the developed models where unpredictable variance is encountered rendering the prediction by these models unreliable, most of the times (*6).*

Recently, advanced machine learning techniques such as Artificial Neural Network (ANN), tree-based algorithms, and other temporal models have been employed in Pavement Management System (PMS) as an alternative to traditional pavement performance prediction models (*3, 7, 8*). These techniques are based on detailed data, require supervised training, and have shown high prediction accuracy. Yet, most of these studies considered limited variables that affect the field performance of AC overlays such as pavement age, climatic conditions, and traffic loading (*3,5,8*), without considering the impact of other key variables such as the pre-treatment pavement condition and the overlay thickness (*9*). Additionally, most of the previous studies predicted the short-term field performance of AC overlays on flexible pavement for a period not exceeding five years (either annually or biannually). Notwithstanding, many researchers have claimed that the developed pavement performance models using data from all or multiple states across the US are questionable in terms of accuracy and should be revised individually based on the roadway network in each state (*3*). Looking to the fact that very few studies investigated the AC overlay performance in Southern State (*5*), none has employed advanced machine learning techniques in their field performance prediction models, nor considered all the variables that significantly impact the AC overlay performance. Thus, there is an urgent need to reflect these two considerations in predicting the long-term field performance of AC overlays in states with hot and humid climates, such as Louisiana.

2. OBJECTIVE

The key objective of this study was to develop a machine-learning based framework that can be used by state agencies in states with hot and humid climate to predict; with a superior accuracy, the long-term field performance (for 11 years) of their AC overlays based on their key project conditions. The two machine learning algorithms used in this study were Random Forest (RF) and CatBoost, and the algorithm yielding the higher accuracy was selected and incorporated into the final framework. The long-term field performance in this study was evaluated in terms of the well-known Pavement Condition Index (PCI). The PCI was developed in the late 1970s by the US Army Corp of Engineers (*10*) and has been widely adopted as a measure of the current condition of the pavement based on the distresses observed on the surface. This index provides a comprehensive indication of the pavement's structural integrity and surface operational condition (i.e., localized roughness and safety) (*11).*

3. BACKGROUND

3.1 Random Forest (RF) Algorithm

The random forest algorithm is based on combining decision tree framework with ensemble learning methods to build multiple trees independent of the original training set. A bagging technique is implemented by generating bootstrap samples and the average of all trees will be finally reported (*12*). Instead of considering all data features at each split, RF uses a random subset of features during splitting in order to reduce the correlation among trees. Such adjustment never builds similar trees which enhances the model accuracy.

3.2 CatBoost Algorithm

CatBoost (or Categorical Boosting) is a new gradient boosting, tree-based ensemble algorithm that outperforms other tree-based algorithms in reducing the gradient boosting biases (*13)*. Trees in CatBoost are grown sequentially such that each tree models the residual errors resulting from the previous trees. It does not use binary substitution of categorical variables like other tree-based algorithms (*14*), instead it handles categorical features by dealing with them during training. Such approach reduces overfitting while handling categorical features and does not compromise accuracy.

Similar to RF, CatBoost is a tree-based algorithm consisting of several decision trees that are combined together to enhance the regression accuracy (*12*). Hence, a general model ($\hat{y}$) of this algorithm can be written as a summation of all scores from all trees for a sample (x). In a single decision tree, the model builds a set of decision rules from the input variables to predict a response variable. The decision rules are called nodes, and split the features space into sub-nodes. These sub-nodes are further split until a specific criterion is met (*12*), where at the end of these structures each one of these terminal sub-nodes will be called leaf.

Assuming (n) observations, (m) input variables, and (C) as the score assigned for each leaf, the general formulation for this structure can be given by **Equation (1)**.

$$f(x)= C_{q(x)}, (q : \mathbb{R}^{m} \rightarrow 1,2,\ldots,t, \quad C \in \mathbb{R}^{m}) \qquad (1)$$

Where $q(x)$ represents the decision rules within a tree that assign a sample of the data to the corresponding leaf index, (t) is the total number of leaves in the tree, and $C_{q(x)}$ represents the score weights assigned to the leaves of the tree (*12*).

Consequently, a general tree-based ensemble model ($\hat{y}$) that consists of multiple trees is presented by **Equation (2)** along with **Equation (1).**

$$\hat{\mathrm{y}}_{\mathrm{i}}(\mathrm{x})=\sum_{\mathrm{T}=1}^{\mathrm{T}} f_{\mathrm{T}}\ (x_i)\ ,\ \left(f_{\mathrm{T}} \in \mathcal{F}\right) \tag{2}$$

Where (T) is the number of trees and ($\mathcal{F}$) is the space of all possible trees. An optimized version of this equation can be given by **Equation (3)**.

$$\mathrm{Obj}\ (\theta)=\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{l}\ \left(\mathrm{y}_{\mathrm{i}},\hat{\mathrm{y}}_{\mathrm{i}}\right)+\sum_{\mathrm{T}=1}^{\mathrm{T}} \Omega\ (f_{\mathrm{T}}) \tag{3}$$

Where l $(y_i, \hat{y}_i)$ is the loss function measurng the difference between the prediction $(\hat{y}_i)$ and target (y_i). The second term is the regularization term that controls the model complexity and prevents overfitting.

3.3 Performance Prediction Models in Previous Studies

In recent years, numerous studies have been conducted to predict the performance of asphalt pavements over its service life with respect to the overall pavement condition. Most performance prediction models were functions of pavement age, surface type, materials used, traffic volume, and climate. In 2006, Kim et al. (*8*) developed a set of asphalt pavement performance prediction models for flexible state and interstate highways. The models were based on data obtained from the Pavement Condition Evaluation System (PACES) database which were collected annually over a period of 15 years (from 1986 to 1999) and analyzed using linear and multiple regression analysis methods. The input variables were PACES Rating, Annual Average Daily Traffic (AADT), and service year. The output of these models was the future PACES rating, which represents the pavement condition. The prediction accuracy in terms of coefficients of determination (R^2) was relatively low; R^2 as low as 0.58 and 0.59 for two years and five years prediction period, respectively.

Similarly, in 2009, Khattak et al. (*5*) developed generalized performance models for multiple distress indices using only two input variables; the pavement surface age (starting from the year when it was last resurfaced) and pavement ’age’ (starting from the year of last reconstruction). The objective of this study was to predict the pavement condition and determine the remaining service life (RSL) for four pavement types and four highway classifications. Researchers have reported that up to 90 percent of the data exhibited ±7.5 percent error between the predicted and observed values.

In 2019, the objective of Yamany et al. (*3*) study was to examine if pavement performance models should be state-specific rather than being based on data from all or multiple states across US. To test this hypothesis, researchers have utilized the Long-Term Pavement Performance (LTPP) condition data for eight states in the US Midwest, namely Indiana, Illinois, Wisconsin, Michigan, Ohio, Minnesota, Iowa, and Missouri, from 1989 to 2016. Combining first the data of all the eight states, three models were developed to estimate the pavement performance of Interstate flexible pavements, namely Fixed-Parameters (FP) regression model, Random Parameters regression (RP) model, and an ANN model. The input variables were the Annual Average Precipitation (AAP), Annual Average Temperature (AAT), Annual Average Freezing Index (AAFI), Annual Average Daily Truck Traffic (AADTT), Equivalent Single-Axle Load (ESAL), and pavement age in years. Then, the empirical data from each of the Midwestern states were individually fed into the three developed to assess their performance at the state level (*3*).

Although the ANN model outperformed FR and RP models for Illinois, Wisconsin, Minnesota, Iowa, and Missouri states, the R^2 values did not exceed 0.72. The RP model was a better fit for the remaining three states, with R^2 value of 0.42, 0.51, and 0.74 for Michigan, Indiana, and Ohio, respectively.

Similar other studies have based their flexible pavement performance predictive models on a limited number of variables such as pavement age, AADT, and weather conditions while neglecting other significant variables such as pre-treatment condition and treatment thickness (*4, 15, 16*).

3.4 Knowledge Gaps in the Literature

According to the reviewed literature, this study addressed several knowledge gaps as follows:

- To the authors' knowledge, most of the previous studies have not considered the pre-treatment pavement condition, overlay thickness, and highway functional class in their prediction. It is well recognized that these variables significantly affect the performance of

AC overlays, particularly the pre-treatment pavement condition and overlay thickness (*5*, *9*). Therefore, in this study, the developed framework considered all the variables that affect the field performance of AC overlays including pre-treatment pavement condition, overlay thickness, and highway functional class.

- Most of the previous studies predicted the field performance of their AC pavements over a period of less than five years, where R^2 values were reported as low as 0.42 (*3*). In this study, the field performance was predicted over a period of 11 years with a superior accuracy (ranging between 0.53 and 0.72).

4. DATA COLLECTION

Data collection in this study was conducted using Louisiana Department of Transportation and Development (LaDOTD) Pavement Management System (PMS) databases. In the LaDOTD PMS, pavement performance data are reported for the period ranging from 1996 to 2019. These data are based on pavement condition measurements that are collected biennially using the Automatic Road Analyzer (ARAN) system that provides a continuous assessment of the road network *(17).*

Collected data are reported every 1/10th of a mile and are analyzed to calculate different distress indices on a scale from zero to 100 (100 being perfect conditions). These indices include the Pavement Condition Index (PCI), Alligator Cracking Index (ALCR), Rutting Index (RUT), Random Cracking Index (RNDM), Roughness Index (RUFF), and Patch Index (PTCH). For flexible pavements, the PCI is calculated as follows *(17):*

PCI = MAX [MIN (RNDM, ALCR, PTCH, RUFF, RUT), {AVG (RNDM, ALCR, PTCH, RUFF, RUT) – 0.85×standard deviation (RNDM, ALCR, PTCH, RUFF, RUT)}] (1)

The Alligator Cracking Index (ALCR) reflects the extent (in terms of cracked area) and severity of alligator cracks existing on the pavement surface and is computed as follows*:*

X= Maximum of 0 and (100-DP_L-DP_M- DP_H) (2)

ALCR= Minimum of 100 and X (3)

Where DP = deduct point due to alligator cracks; and subscripts L, M, and H refer to the low, medium, and high severity of the cracks, respectively.

The Roughness Index (RUFF) reflects the irregularities in the pavement surface and is expressed on a scale from zero to 100 with 100 representing the case of a smooth pavement. It is related to the IRI using the following empirical equation *(17):*

IRI (in/mile) = (100 - RI) × 5 + 50 (4)

The Rutting Index (RUT) reflects the average rutting depth (R_AVG) in the pavement surface, and is expressed in a scale from 0 to 100 with 100 representing the case with no rutting. This index is calculated as follows:

If (R_AVG>=0 mm and R_AVG<3.1 mm), then RUT=100 (5)

If (R_AVG>=3.1 mm and R_AVG<35 mm), then RUT=-80× (R_AVG [in inch]) +110 (6)

If (R_AVG>=35 mm), then RUT=0 (7)

In this study, a total of 50 AC overlay sections were identified from LaDOTD databases. To provide an accurate prediction, the analysis of these sections was conducted for every log-mile (0.1 mile), which was considered as a single data point. This resulted in a total of 892 log-miles (data points). For every log-mile, the following was reported:

1. Six overlay age values (A): A_1, A_3, A_5, A_7, A_9, and A_{11} where A_1 represents one year after the overlay application, A_3 represents three years after the overlay application, and similarly for A_5, A_7, A_9, and $A_{11.}$
2. Seven measured PCI values: $MPCI^-$ (before AC overlay application), in addition to $MPCI_1$, $MPCI_3$, $MPCI_5$, $MPCI_7$, $MPCI_9$, and $MPCI_{11}$ which correspond to A_1, A_3, A_5, A_7, A_9, and A_{11} respectively.
3. Six Annual cumulative Truck Traffic (TT): TT_1, TT_3, TT_5, TT_7, TT_9, and TT_{11} which correspond to A_1, A_3, A_5, A_7, A_9, and A_{11}, respectively. In this study, TT is defined as the sum of the annual volume of trucks observed on a given road section up to a specific year.
4. Six Annual cumulative rainfall (R): R_1, R_3, R_5, R_7, R_9, and R_{11} which correspond to A_1, A_3, A_5, A_7, A_9, and A_{11}, respectively. In this study, R is defined as the sum of the annual precipitation observed over a road section in a specific region up to a specific year.
5. Six mean annual temperature values (T): T_1, T_3, T_5, T_7, T_9, and T_{11} which correspond to A_1, A_3, A_5, A_7, A_9, and A_{11}, respectively. In this study T is defined as the average temperature observed over a road section in a specific region over the entire year.
6. Highway function classification (C): Either principal arterial, minor arterial, or major collector. Label encoding method was applied to encode this categorical variable where values of 1, 2, and 3 were assigned for levels principal arterial, minor arterial, and major collector, respectively.
7. Overlay thickness (OT) in inches.

6.CORRELATION

A correlation matrix showing the linear relationship between the different sets of input variables and the final PCI measurement at age 11 ($MPCI_{11}$) was examined, see Table **1**. In general, the correlation coefficient ranges between −1.0 and 1.0, where a value of 1.0 means a perfect, increasing, linear relationship and −1.0 means a perfect, decreasing, linear relationship. As shown in Table **1**, PCI^- had the highest correlation to $MPCI_{11}$ (correlation coefficient of 0.83) which supports the importance of considering the pre-treatment pavement condition in predicting the future PCI of AC overlays. R_{11} had the second highest correlation to $MPCI_{11}$ (correlation coefficient of -0.46) which; as expected, indicates the significant impact of rainfall on the future performance of AC overlays in states with hot and humid climate.

Table 1 Correlation Matrix between input variables and MPCI11

	C	OT	A	TT_{11}	R_{11}	T_{11}	PCI^-	$MPCI_{11}$
C	1.0	-0.04	0.0	-0.33	0.27	0.18	-0.10	-0.11
OT	-0.04	1.0	0.0	-0.04	0.02	0.13	0.11	0.10
A	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
TT_{11}	-0.33	-0.04	0.0	1.0	0.22	0.14	-0.28	-0.15
R_{11}	0.27	0.02	0.0	0.22	1.0	0.21	-0.59	-0.46
T_{11}	0.18	0.13	0.0	0.14	0.21	1.0	-0.05	-0.06
PCI^-	-0.10	0.11	0.0	-0.28	-0.59	-0.05	1.0	0.83
$MPCI_{11}$	-0.11	0.10	0.0	-0.15	-0.46	-0.06	0.83	1.0

6. MODEL DEVELOPMENT

6.1 Model Overview

The Random Forest (RF) algorithm was used to develop six different models (A to F) that could be used sequentially to predict the PCI at age 1 ($PPCI_1$), PCI at age 3 ($PPCI_3$), PCI at age 5 ($PPCI_5$), PCI at age 7 ($PPCI_7$), PCI at age 9 ($PPCI_9$), and PCI at age 11 ($PPCI_{11}$), respectively based on the aforementioned collected variables (C, OT, A, TT, R, T, and PCI^-). Figure 1 shows a schematic of the inputs and output for each of the six developed models. Each of these models was trained using 80% of the collected data (713 points) and was then tested using the remaining 20% of the data (179 points). Similarly, the CatBoost algorithm was used to develop six other models (L to Q) using the same concept in Figure 1. Based on Figure 1, the general formulation for the six models for RF and CatBoost algorithms is represented in the following equations:

Models A and L

$$PPCI_1 = f\ (C, OT, A_1, TT_1, R_1, T_1, PCI^-) \tag{5}$$

Models B and M

$$PPCI_3 = f\ (C, OT, A_3, TT_3, R_3, T_3, PPCI_1) \tag{6}$$

Models C and N

$$PPCI_5 = f\ (C, OT, A_5, TT_5, R_5, T_5, PPCI_3) \tag{7}$$

Models D and O

$$PPCI_7 = f\ (C, OT, A_7, TT_7, R_7, T_7, PPCI_5) \tag{8}$$

Models E and P

$$PPCI_9 = f\ (C, OT, A_9, TT_9, R_9, T_9, PPCI_7) \tag{9}$$

Models F and Q

$$PPCI_{11} = f\ (C, OT, A_{11}, TT_{11}, R_{11}, T_{11}, PPCI_9) \tag{10}$$

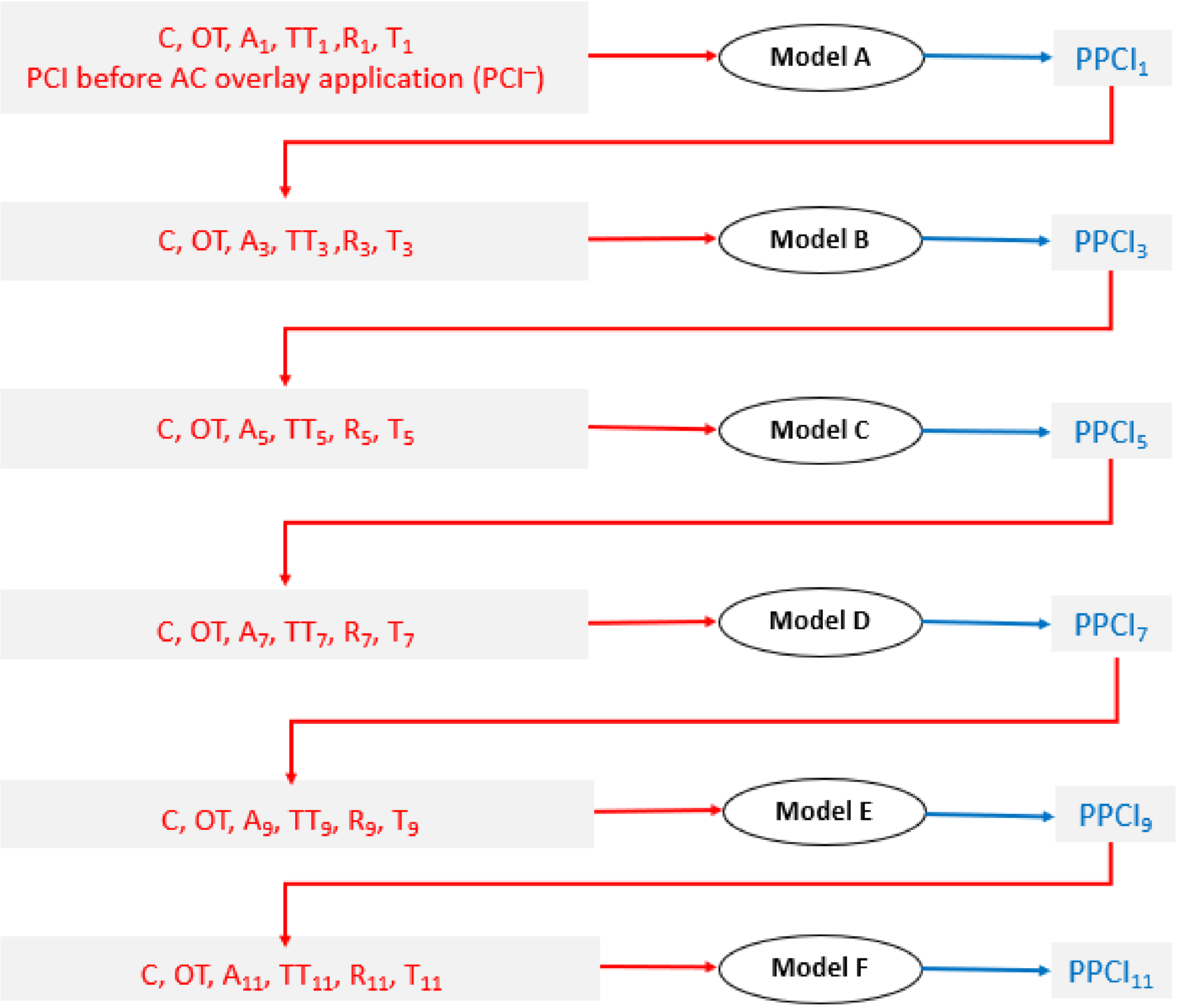


Figure 1 Schematic of the inputs and output of each of the six developed models

6.2 Model Training

Both RF and CatBoost algorithms have hyper-parameters that need to be optimized during the training phase to ensure optimized models' accuracy. The RF algorithm requires tuning of the following hyper-parameters (*12*):

a) Subset of features (S): indicates the size of the variable subset randomly sampled from the original set of variables while developing each RF tree.

b) Number of Trees (T): is the total number of trees in the model that would be averaged.

As for CatBoost algorithm, it requires tuning of the following hyper-parameters (*12*):

a) Maximum tree depth (D): is the maximum number of successive nodes/split in the tree.

b) T: as described above.

c) Learning rate (L): is the learning rate which shrinks the contribution of each successive tree by the value of L, therefore, overcoming any overfitting problem.

To optimize these hyper-parameters, two combined techniques were used: grid search and ten-fold cross validation. Generally, grid search examines all possible combinations of hyper-parameter values within a defined space to identify the optimal combination. For all the RF and CatBoost models, the different parameter spaces were defined as $S \in [1,2, 3, \ldots, 45]$, $D \in [1, 2, 3, \ldots, 10]$, $T\in [10, 20, 30, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]$, and $L \in [0.005, 0.01, 0.05, 0.1, 0.3, 0.5]$. The grid search was guided by a ten-fold cross validation technique in which the data were divided into 10 subsets. Then, the model training was performed using nine subsets and validation was done using the remaining subset. This was repeated 10 times by changing the validation subset. For each trial, the R^2 was obtained, and the average R^2 value was finally obtained for the ten trials to evaluate the model performance. Table *2* presents the optimized hyper-parameters and the corresponding R^2 and RMSE for all the RF and CatBoost models.

Table 2 Optimal Hyper-parameters and corresponding accuracy for RF and CatBoost Models

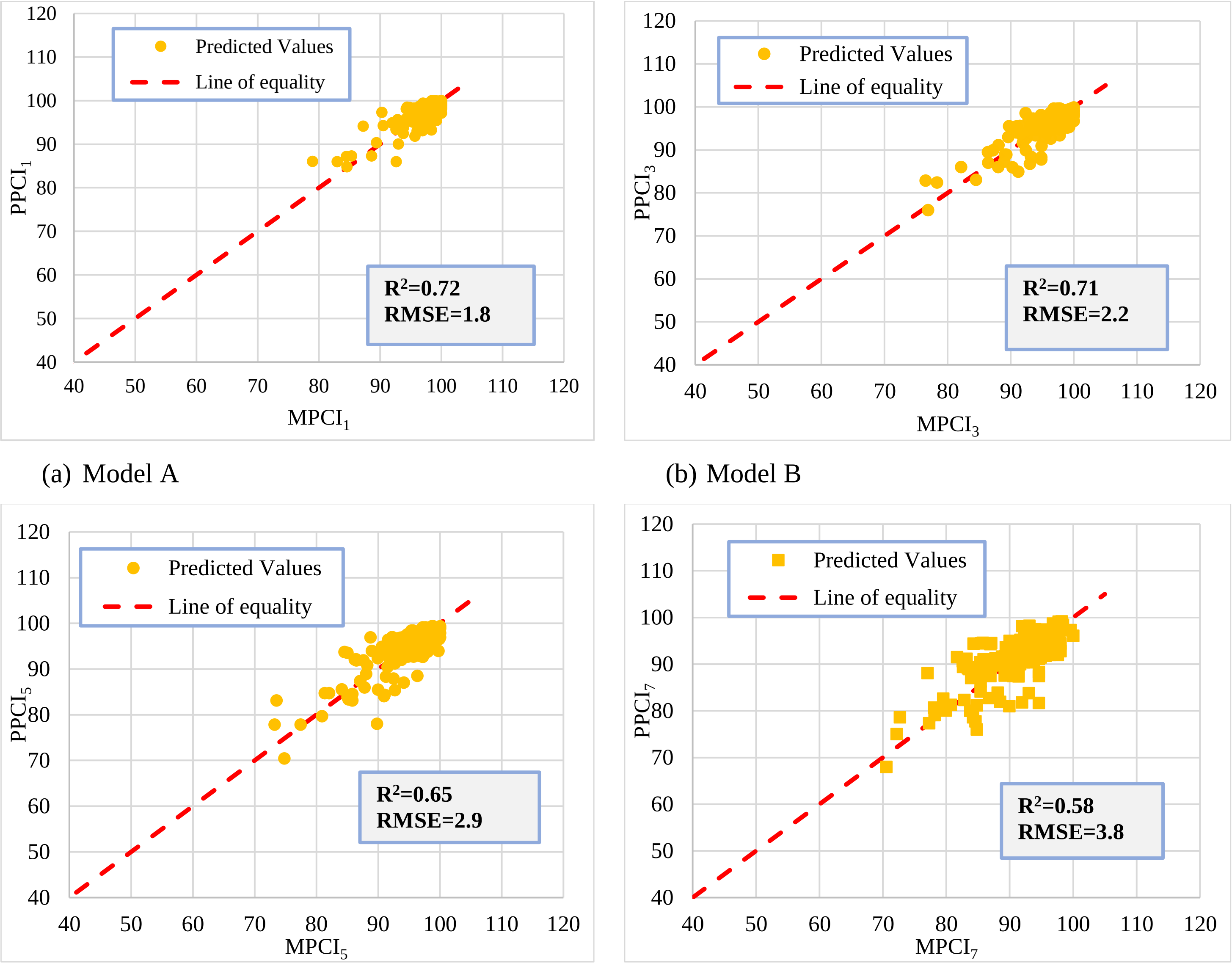
<table>
<tr><th colspan="6">RANDOM FOREST</th></tr>
<tr><th>Models</th><th colspan="2">Subset of features (S)</th><th>Number of trees (T)</th><th>R^2</th><th>RMSE</th></tr>
<tr><td>A</td><td colspan="2">1</td><td>700</td><td>0.91</td><td>1.0</td></tr>
<tr><td>B</td><td colspan="2">3</td><td>400</td><td>0.90</td><td>1.3</td></tr>
<tr><td>C</td><td colspan="2">10</td><td>900</td><td>0.92</td><td>1.4</td></tr>
<tr><td>D</td><td colspan="2">2</td><td>1000</td><td>0.88</td><td>2.2</td></tr>
<tr><td>E</td><td colspan="2">1</td><td>600</td><td>0.86</td><td>2.8</td></tr>
<tr><td>F</td><td colspan="2">2</td><td>1000</td><td>0.87</td><td>3.1</td></tr>
<tr><th colspan="6">CATBOOST</th></tr>
<tr><th>Models</th><th>Maximum tree depth (D)</th><th>Number of trees (T)</th><th>Learning rate (L)</th><th>R^2</th><th>RMSE</th></tr>
<tr><td>L</td><td>3</td><td>400</td><td>0.1</td><td>0.87</td><td>1.2</td></tr>
<tr><td>M</td><td>3</td><td>400</td><td>0.1</td><td>0.84</td><td>1.7</td></tr>
<tr><td>N</td><td>3</td><td>400</td><td>0.1</td><td>0.77</td><td>2.7</td></tr>
<tr><td>O</td><td>3</td><td>400</td><td>0.1</td><td>0.76</td><td>3.1</td></tr>
<tr><td>P</td><td>3</td><td>400</td><td>0.1</td><td>0.75</td><td>3.8</td></tr>
<tr><td>Q</td><td>1</td><td>1000</td><td>0.01</td><td>0.67</td><td>4.9</td></tr>
</table>

As shown in Table *2*, based on the training data, all the RF models had higher accuracy (higher R^2 and lower RMSE) than the corresponding CatBoost models. Therefore, only the RF models will be considered throughout the remaining of this study.

As expected, the general trend in Table *2* shows that the accuracy decreases from Model A to F since the prediction horizon increases from 1 year to 11 years. The RF algorithm predicted the PCI of AC overlays after 1 year (Model A) with an R^2 of 0.91 and RMSE of 1.0, and predicted the PCI of AC overlays after 11 year (Model F) with the an R^2 of 0.87 and RMSE of only 3.1. Comparing these values to the general R^2 reported in the literature which ranged between 0.42 and 0.74 for a prediction period up to five years (*3*, *8*), one can conclude that the RF algorithm predicted the PCI values with a superior level of accuracy.

6.3 Model Testing

This section presents the performance of the six RF models (Models A to F) using the testing data. Figure 2 (a - f) show the relation between the measured and predicted PCI values for Models A to F. As shown, based on the testing data, the models predicted the PCI with a relatively high accuracy in the first five years (R^2 ranging between 0.72 and 0.65 and RMSE ranging between 1.8 and 2.9) as compared to the values reported in the literature, which had an R^2 as low as 0.42 in the first five year (*3*). After seven, nine, and eleven years, the prediction accuracy decreased with R^2 of 0.58, 0.58, and 0.53, respectively and RMSE of 3.8, 4.6, and 5.9, respectively. Yet, these values are still reasonable compared to the values reported in the literature for a prediction period exceeding five years. It should be noted this data was not considered in the model training, and thus would reflect the models' accuracy. Based on the analysis in this section, it was concluded that the developed RF models can be used by state agencies to predict the future PCI of their AC overlays in the first five years with a high reliability and in the following six years with a relatively lower reliability.

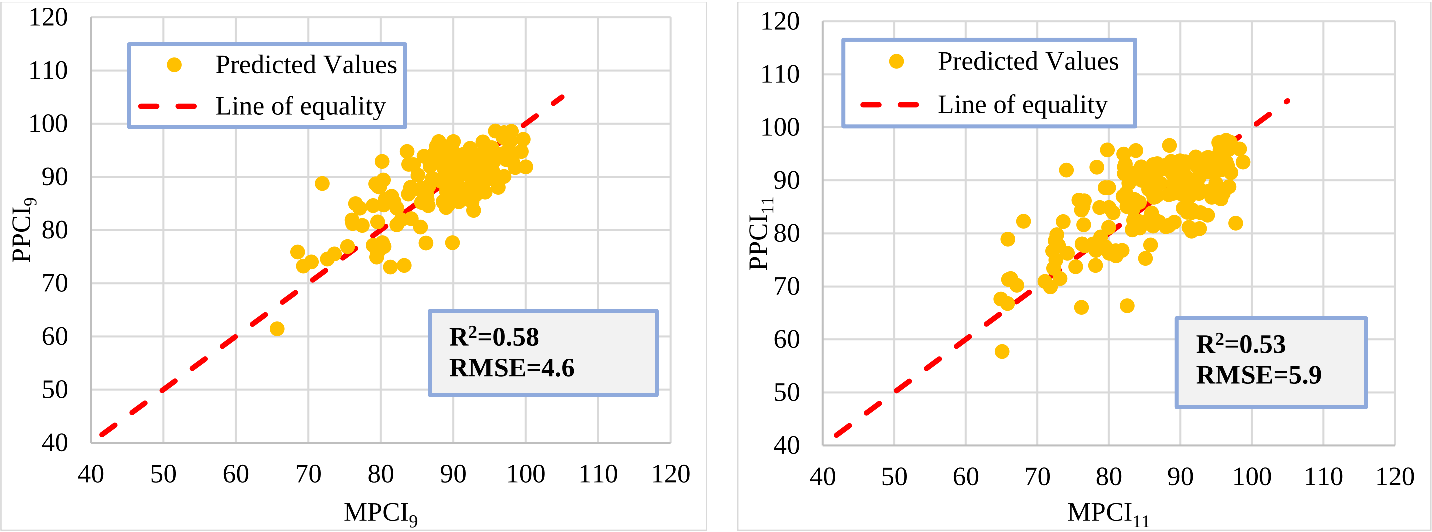


(a) Model A

(b) Model B

(c) Model C

(d) Model D



(e) Model E

(f) Model F

Figure 2 Performance of Random Forest algorithm using the testing data

6.4 Relative Importance of Model Input Variables

The Relative Importance (*18*) is a statistical measure defined as the percentage contribution of each input variable to the model when these variables are dependent and not directly manipulated. The Relative Importance of a variable is calculated as the total gain from this variable across all trees and normalized such that all variables add up to 100 (*19*). The higher the value, the greater the variable's contribution to the model. Figure 3 shows the Relative Importance of each input variable to the six RF models (A to F).

As illustrated in Figure 3, the most important variable in all the RF models (A to F) was the predicted PCI at the pervious time step (or the pre-treatment pavement condition PCI^- for Model A). This agrees with the results of the measured PCI values in Table **1** which showed that the measured PCI has the highest correlation to the pre-treatment pavement conditions. This could explain the low accuracy of the PCI prediction models presented in the literature that did not consider the pre-treatment pavement conditions. Figure 3 also shows that the traffic level, rainfall, and the mean temperature are important variables in the PCI prediction. This is expected as they accelerate the pavement deterioration, which is in agreement with previous studies (*4*,*15*,*16*).

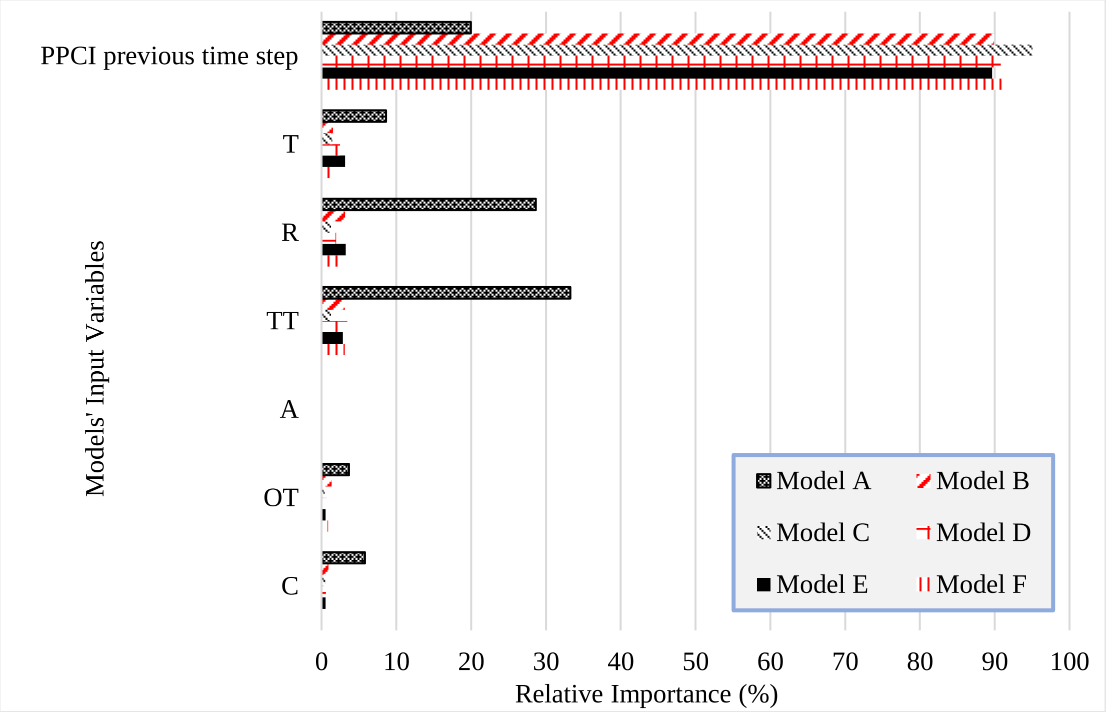


Figure 3 Relative Importance percentage of the input variables

7. ILLUSTRATIVE APPLICATION OF THE DEVELOPED MODELS

After the application of AC overlays, state agencies are interested in predicting the future PCI values of the applied AC overlays. This is essential for determining the expected service life of the applied AC overlays to plan for future maintenance and treatment activities and economically allocate the corresponding funds. The implementation of the RF models developed in this study (Models A to F) are expected to assist in this decision-making process as outlined in the following steps (with the help of Figure 4 and Table *3* which present a numerical example for one of the log-miles included in the testing data of this study):

Step 1: Collect the Pre-treatment Pavement Condition Index

Collect the last PCI measured before the application of the AC overlay (PCI^-). This could be readily obtained from the PMS databases. In the example in Table *3*, this value was 73.17.

Step 2: Collect the other Variables Biannually over the 11-year Period

Collect all the other 6 inputs in Table *3* (C, OT, A, TT, R, and T) at 1, 3, 5, 7, 9, and 11 years. It should be noted that the highway function classification (C) and overlay thickness (OT) would be constant throughout the time intervals and could be easily obtained.

Step 3: Use the Developed Six Models Sequentially

In this step, the user will input all the input data at year 1 into Model A to predict the PPCI at year 1 ($PPCI_1$=93.81 in Table *3*). After that, the user will use all the input data at year 3 into Model B along with the $PPCI_1$ to predict $PPCI_3$ ($PPCI_3$=88.88 in Table *3*). This process will be recursively applied, i.e. from Model C to Model F, until the $PPCI_{11}$ is predicted. Figure 4 presents the measured PCI (as collected from LaDOTD PMS) as compared to the predicted PCI using Models A to F for the example presented in Table *3*.

Table 3 Example Results

Variable Type	Variables	Reported Values					
		Model A	Model B	Model C	Model D	Model E	Model F
Input	C	Minor Arterial	Minor Arterial	Minor Arterial	Minor Arterial	Minor Arterial	Minor Arterial
	OT	4	4	4	4	4	4
	A	1	3	5	7	9	11
	TT	1,350,135	3,587,220	4,467,600	5,488,140	6,889,740	8,243,160
	R	49.66	173.12	283.65	410.42	500.45	631.2
	T	67.3	68.3	68.1	67.7	68.0	66.9
	Input PCI	73.17	93.81	88.88	84.74	80.73	75.48
Output	PPCI	93.81	88.88	84.74	80.73	75.48	70.23
Actual	Actual PCI	93.8	89.08	81.35	78.05	73.6	67.14

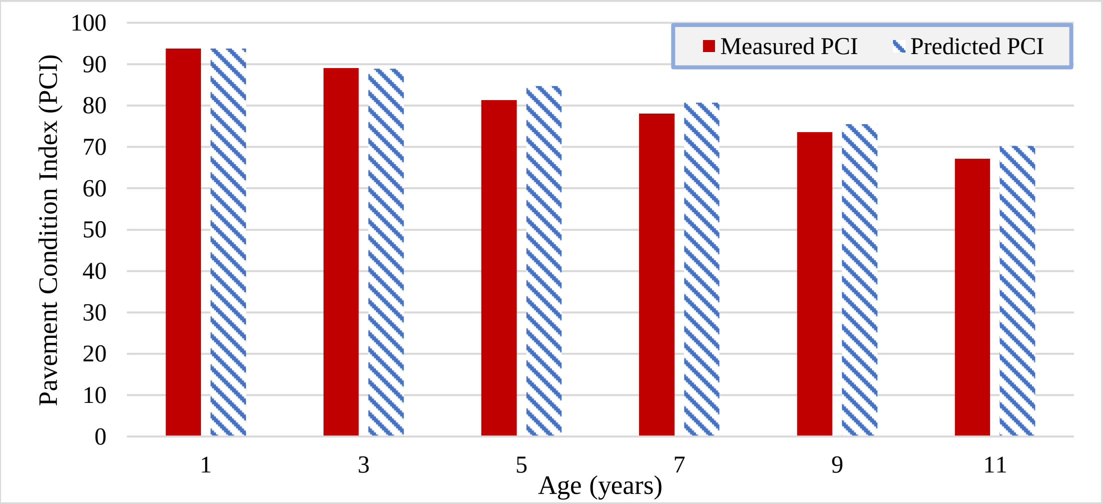


Figure 4 Measured (actual) and predicted PCI for the example in Table 3

8. SUMMARY AND CONCLUSIONS

AC overlays are one of the most common rehabilitation techniques used to restore pavement conditions. It is crucial for highway agencies to accurately monitor the deterioration of these overlays over time and determine the contribution of the important factors to AC overlay performance. Yet, few studies to date have considered machine-learning for modeling the field performance of AC overlays on flexible pavements, particularly in hot and humid climates, while considering all important influential variables on AC overlay performance.

Hence, the objective of this study was to recommend a machine-learning based framework for state agencies in hot and humid climate regions to predict the long-term field performance (for 11 years) of their AC overlays based on their key project conditions. The long-term field performance was the PCI, and the algorithms utilized for this prediction were RF and CatBoost. A total of 892 log miles of AC overlay data were collected from LaDOTD PMS database and analyzed. These 892 log miles were included in 50 asphalt overlay projects on existing flexible pavement. Based on the data, the research team developed six models to be employed sequentially to predict the PCI at age 11 ($PPCI_{11}$) based only on highway function classification (C), overlay thickness (OT), overlay age (A), annual cumulative truck traffic (TT), annual cumulative rainfall (R), mean annual temperature (T), and PCI before overlay application (PCI^-). At the data training phase, RF outperformed CatBoost in PCI prediction, with R^2 ranging from 0.91 at Age 1 to an R^2 of 0.87 at Age 11. At the testing phase, the developed RF models have showed a relatively high accuracy in the first five years prediction period (R^2 ranging between 0.72 and 0.65 and RMSE ranging between 1.8 and 2.9), and this prediction accuracy decreased to lower values after seven, nine, and eleven years (R^2 of 0.58, 0.58, and 0.53, respectively and RMSE of 3.8, 4.6, and 5.9, respectively). The resulting RF models are expected to assist transportation agencies throughout the South-Central US in predicting the PCI of their AC overlays with high accuracy based on one PCI measurement at the pre-treatment stage and other project condition inputs.

Furthermore, from the relative importance evaluation of input variables, the predicted PCI at the pervious time step, the traffic level, rainfall, and the mean temperature showed to have significant contribution on the performance of AC overlays. Thus, it is important to include these variables in future AC overlay prediction models.

ACKNOWLEDGMENTS

The financial support of the Transportation Consortium of South-Central States (Tran-SET) is greatly appreciated (Tran-SET Project 21PLSU15and Grant Number 69A3551747106).

REFERENCES

1. Chang, J. R., D. H. Chen, and C. T. Hung. (2005). Selecting Preventive Maintenance Treatments in Texas: Using the Technique for Order Preference by Similarity to the Ideal Solution for Specific Pavement Study-3 Sites. Transportation Research Record: Journal of the Transportation Research Board, Vol. 1933, pp. 62–71

2. Prozzi, J.A., and Madanat, S.M. (2004). Development of Pavement Performance Models by Combining Experimental and Field Data. Journal of Infrastructure Systems, 10(1):9-22.

3. Yamany, M.S., Saeed, T., Volovski, M., and Ahmed, A. (2019). Characterizing the Performance of Interstate Flexible Pavements using Artificial Neural Networks and Random Parameters Regression. J. Infrastruct. Syst., 2020, 26 (2): 04020010

4. Lou, Z., Gunaratne, M., Lu, J., and Dietrich, B. (2001). Application of Neural Network Model to Forecast Short-Term Pavement Crack Condition: Florida Case Study. Journal of Infrastructure Systems, 7(4): 166-171.

5. Khattak, M.J., Baladi, G. Y., and Sun, X. (2009). Development of Index Based Pavement Models for Pavement Management System (PMS) of LADOTD. Federal Highway Administration FHWA/LA.08/460. Louisiana Transportation Research Center, Baton Rouge.

6. Umali, J., and Erniel B. (2014). "Nonparametric principal components regression." Communications in Statistics-Simulation and Computation 43.7 10 (2014): 1797-1810.

7. Sundin, S. & Bradan-Ledoux, C. (2002). Artificial Intelligence- Based Decision Support Technologies in Pavement Management. Computer-Aided Civil and Infrastructure Engineering, 16 (2), 143-157, https://doi.org/10.1111/0885-9507.00220

8. Kim, S. H., and Kim, N. (2006). Development of Performance Prediction Models in Flexible Pavement using Regression Analysis Method. KSCE J. Civ. Eng. 10 (2): 91–96. https://doi.org/10.1007/BF02823926.

9. Saeed, T. U., Qiao, Y., Chen, S., Alqadhi, S., Zhang, Z., Labi, S., and K. C. Sinha (2017). Effects of Bridge Surface and Pavement Maintenance Activities on Asset Rating. Joint Transportation Research Program, Publication No. FHWA/IN/JTRP-2017/19. West Lafayette, IN: Purdue Univ.

10. Shahin, M. Y., Darter, M.I., and Kohn, S.D. (1980). “Condition Evaluation of Jointed Concrete Airfield Pavement.” Transportation Engineering Journal of ASCE, 106(4), pp. 381–399, American Society of Civil Engineers, Reston, VA

11. Arhin, S.A., and Noel, E.C. (2014). Predicting Pavement Condition Index from International Roughness Index in Washington, DC. Howard University Transportation Research Center. DDOT-RDT-14-03,

12. Mousa, S. R., Bakhit, P. R., Osman, O. A., & Ishak, S. (2018). A Comparative Analysis of Tree-Based Ensemble Methods for Detecting Imminent Lane Change Maneuvers in Connected Vehicle Environments. Transportation Research Record, 2672(42), 268-279.

13. Mousa, M. and Elseifi M. (2019). Development of Tree-Based Algorithm for Prediction of Field Performance of Asphalt Concrete Overlays. Journal of Transportation Engineering, Part B: Pavements /Volume 145 Issue 2- June .

14. Ray, S. (2017). CatBoost: A Machine Learning Library to Handle Categorical (CAT) Data Automatically. Analytics Vidhya, 14 Aug., https://www.analyticsvidhya.com/blog/2017/08/catboost-automated-categorical-data/. Accessed on June 1, 2022.

15. Amin, S. R., and Amador-Jiménez, L. E. (2016). Backpropagation Neural Network to Estimate Pavement Performance: Dealing with Measurement Errors. Road Mater. Pavement Des. 18 (5): 1218–1238. https://doi.org /10.1080/14680629.2016.1202129.

16. Gulen, S., Zhu, K., Weaver, J., Shan, J., and Flora, W. (2001). Development of Improved Pavement Performance Prediction Models for the Indiana Pavement Management System. Publication FHWA/IN/JTRP-2001/17. Joint Transportation Research Program, Indiana Department of Transportation and Purdue University, West Lafayette, Indiana

17. Mousa, M., Elseifi, M. A., Bashar, M. Z., Zhang, Z., and Gaspard, K. Short and long-term field performances and optimal timing of chip seal in hot and humid climates. Transportation Research Record, 2674(1), 2020, pp. 33-43

18. Mousa, M., Mousa, S., Hassan, M., and Carlson, P. (2021). Predicting the Retroreflectivity Degradation of Waterborne Paint Pavement Markings using Advanced Machine Learning

techniques. Transportation Research Record: Journal of the Transportation Research Board, 2675 (7), DOI: 10.1177/03611981211002844

19. Chen, T., and Guestrin, C.. (2016). Xgboost: A scalable tree boosting system.In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 785-794.