

1 **Modelling of Pavement Performance Evolution Considering Uncertainty and**
2 **Interpretability: A Machine Learning Based Framework**

3 Linyi Yao^a, Zhen Leng^{a,*}, Jiwang Jiang^a, Fujian Ni^b

4 ^a *Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University,*
5 *Hung Hom, Kowloon, Hong Kong*

6 ^b *Department of Highway and Railway Engineering, School of Transportation, Southeast University,*
7 *Nanjing, Jiangsu, China*

8 *corresponding author

9 Email: zhen.leng@polyu.edu.hk

1 **Modelling of Pavement Performance Evolution Considering Uncertainty** 2 **and Interpretability: A Machine Learning Based Framework**

3 Machine learning (ML) based pavement performance models have gained increasing
4 popularity in recent years due to their strong power in modelling complex relationships.
5 However, the insufficiency of a feature selection process prior to model construction,
6 the difficulty in explaining the black box models, and the lack of uncertainty
7 consideration all impeded the application of the produced models in real world. To fill
8 these gaps, this study aims to develop a new framework to model the pavement
9 performance evolution based on the state-of-the-art ML techniques, including the
10 BorutaShap method for feature selection, the Bayesian neural network (BNN) for
11 model development and uncertainty quantification, and the SHapley Additive
12 exPlanations (SHAP) approach for model interpretation. A case study of predicting the
13 pavement transverse cracking was conducted. The two generated BNN models yielded
14 relatively accurate predictions with the R-square of 0.86 and 0.79 for unmaintained and
15 maintained segments, respectively. Poor data quality was found to be the dominant
16 source of uncertainty. The model interpretation also provided some insight into the
17 underlying influential mechanism of various factors. The framework was expected to
18 enable the decision-makers to build more reliable and informative pavement
19 performance models that could be integrated into the pavement management tools.

20 *Keywords:* Pavement performance model; Machine learning; Feature selection;
21 Uncertainty quantification; Model interpretation.

22

23 **1 Introduction**

24 Pavement performance model is one of the main building blocks in modern pavement
25 management systems (PMS). An accurate pavement performance model can help
26 significantly reduce the long-term maintenance cost since it allows agencies to proactively
27 conduct maintenance and rehabilitation (M&R) on roadway infrastructures (Rahman *et al.*,
28 2017; Yao *et al.*, 2020). However, modelling of pavement performance is a complex and
29 challenging task, as pavement deterioration is influenced by many factors.

30 The available pavement performance models can be commonly classified into three
31 types, namely empirical, mechanistic, and mechanistic-empirical models (Yehia and Sweil,

1 2020). Mechanistic models rely heavily on the expected theoretical behaviour of the materials.
2 However, for pavement materials that suffer from large variations, these models sometimes
3 may produce unexpected results (Choi *et al.*, 2004). In contrast, empirical models are
4 formulated on the basis of statistical correlations between pavement performance and various
5 explanatory variables, which include both simple regression models, such as linear,
6 exponential, and sigmoid, and complicated machine learning (ML) models. Simple regression
7 models are frequently criticized for their low accuracy, which is mainly caused by the
8 inadequate model formulation. Although ML models have demonstrated their power in
9 modelling pavement deterioration, there are still some problems that have not been properly
10 addressed which would be discussed in more details later. Mechanistic-empirical (ME)
11 models adopt mechanistic approaches to compute the critical pavement responses and
12 empirical methods to correlate the responses to pavement failure (Li *et al.*, 2011). But some
13 researchers have proved that ME models tend to perform worse than ML models (El-Hakim
14 and El-Badawy, 2013).

15 It should be recognized that no single prediction method applies to all scenarios.
16 Building a systematic framework for modelling pavement performance evolution is somehow
17 more important. This study mainly focusses on the ML based pavement performance models,
18 considering that ML models can better capture the complex nonlinear relationships among
19 various variables and are therefore more promising for current and future pavement
20 management practice. Previous studies have extensively explored the potential of applying
21 ML techniques to model pavement deterioration (Gong *et al.* 2018; Tabatabaee *et al.* 2013).
22 However, as explained in the following paragraphs, the existing ML models still have some
23 obvious limitations despite the efforts to improve the predictive abilities of the models by
24 means of sophisticated methods.

25 Firstly, the selection of explanatory variables according to their importance level to
26 the response variable was rarely considered. Researchers may have reached a consensus on
27 the variable types that should be included in a pavement performance model. When it comes
28 to the specific variables, it depends mostly on the modelers' experience and the available data.
29 This may result in either the inclusion of irrelevant and redundant features or the exclusion of
30 important features. The former will increase the computational time, reduce the model
31 accuracy, and impede the understanding of the model (Cai *et al.*, 2018), while the latter may
32 make the model unable to fully capture the variations in the dependent variables. For example,

1 equivalent single axle load (ESAL) is one of the most commonly used traffic variables in
2 previous pavement performance models (Karlaftis and Badr, 2015; Mazari and Rodriguez,
3 2016; Choi and Do, 2020). However, the validity of the expression of ESAL has always been
4 controversial among pavement researchers (Prozzi and Madanat, 2004; Guler and Madanat,
5 2011). Thus, the use of multiple traffic variables including axle load spectrum information
6 may be more beneficial to fully characterize the traffic conditions and predict the progression
7 of pavement distresses. On the other hand, the rapidly increasing volume and complexity of
8 the data in today's PMS allow the formulation of various features. Hence, without a complete
9 assessment, incorporating certain features that are deemed sustainable may actually generate
10 counterproductive results. Several recent studies investigated the importance of attributes to
11 pavement performance (Piryonesi and El-Diraby, 2020; Gong *et al.*, 2019; Zeiada *et al.*,
12 2020). However, attempts to incorporate such importance evaluation results into the selection
13 of model inputs are still limited. Meanwhile, the existing literature differed from this study in
14 that most of them tried to find a possible compact subset of features that minimizes the model
15 error (El-Diraby, 2020; Zeiada *et al.*, 2019; Zeiada *et al.*, 2020) or selected a certain number
16 of features according to the order of feature importance (Roberts *et al.*, 2021; Yao *et al.*,
17 2019). In contrast, this study aims to capture more features relevant to the output variable
18 with a so-called all-relevant feature selection method, which is beneficial to understanding
19 the mechanisms of the problems instead of merely building a black box model (Kursa and
20 Rudnicki, 2010).

21 Secondly, pavement deterioration has inherent uncertainty due to measurement error,
22 data processing, and other possible reasons (Yehia and Swei, 2020; Amin and Amador-
23 Jiménez, 2017). This motivates many researchers to incorporate uncertainty consideration
24 into their modelling framework, which helps evaluate the reliability of prediction results. A
25 large number of existing studies have estimated the uncertainties in performance models by
26 assessing the measurement errors in pavement condition data, considering the discrete state
27 space and Markov chain model (Osorio-Lird *et al.*, 2018; Kobayashi *et al.*, 2012). Only
28 limited research extended to the continuous state space. However, these studies tend to focus
29 on either data-driven (aleatoric) or model-driven (epistemic) uncertainty alone (Hong and
30 Prozzi, 2006; Guo *et al.*, 2020), resulting in insufficient estimation of uncertainties from
31 different sources in pavement performance models. To the best of the authors' knowledge,
32 none of the previous ML-based pavement performance models have quantified both the
33 aleatoric and epistemic uncertainties in the continuous state space.

1 Finally, it is worth noting that explanatory modelling is highly desired by practitioners
2 because they need to understand the effects of specific features on the model output for real-
3 world applications. Otherwise, the validity of the decisions derived from the prediction of the
4 ML models cannot be guaranteed. In the existing literature that uses ML to model pavement
5 deterioration, researchers tend to focus on model accuracy only (Dong *et al.*, 2019; Kırbaş
6 and Karaşahin, 2016; Hossain *et al.*, 2019), or utilize the simplest one-at-a-time (OAT)
7 approach to estimate the feature importance (Hossain *et al.*, 2019; Yao *et al.*, 2019; Kargah-
8 Ostadi and Stoffels, 2015). OAT means varying one input variable at a time and observing
9 the maximum range of variation in output variable. However, this method cannot reflect the
10 interactions among input variables, which is statistically incorrect when applied to nonlinear
11 models (Saltelli, 1999; Saltelli *et al.*, 2019). Besides, even though these ML-based models
12 involve some interpretation of the model results, they often stop at evaluating the feature
13 importance or sensitivity without delving into the feature effect (Gong *et al.*, 2019; Yao *et al.*,
14 2019; Tabatabaee *et al.*, 2013).

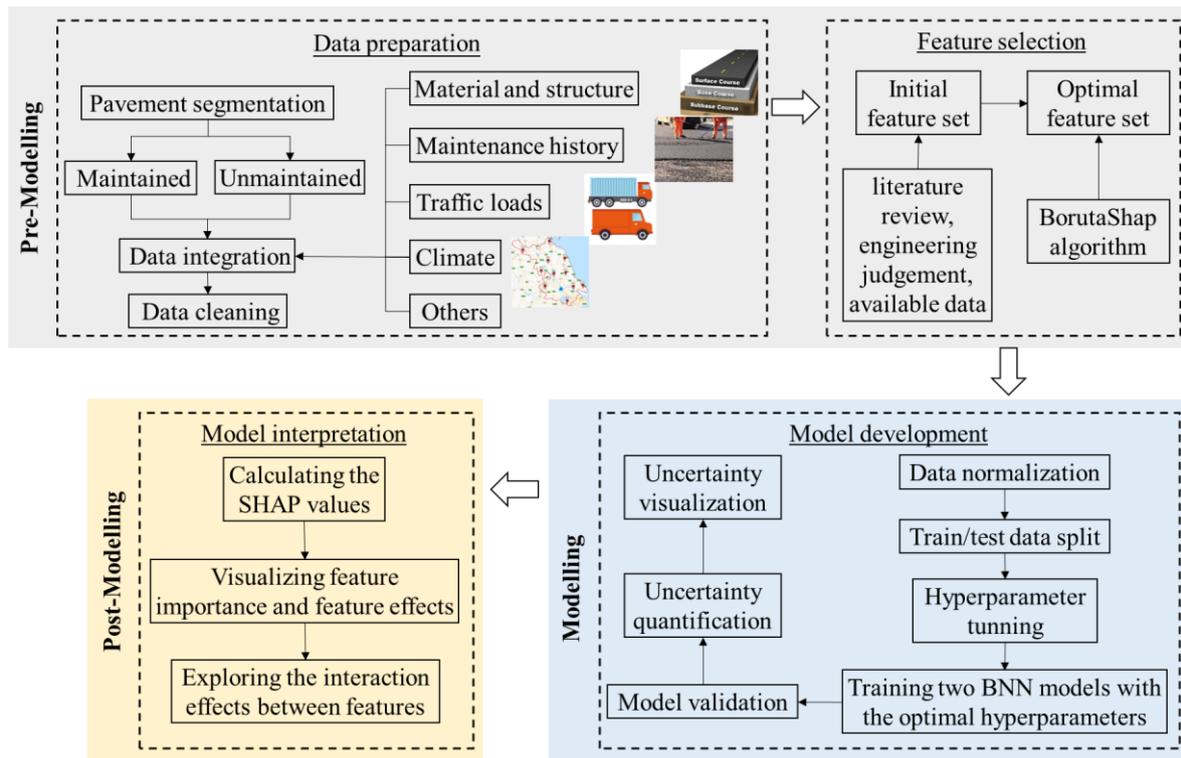
15 To address the aforementioned limitations, this research aims to establish a pavement
16 performance modelling framework based on state-of-the-art ML techniques while taking into
17 account the feature selection, uncertainty estimation and model interpretation problems. The
18 proposed modelling framework is expected to help transport agencies make better use of the
19 increasing available pavement condition data. It will contribute to improved understanding of
20 the uncertainties underlying the pavement performance prediction tools, leading to more
21 confident projection of the future road infrastructure conditions. Moreover, by introducing
22 advanced ML model interpretation techniques, the established models can help explore the
23 pavement deterioration mechanisms under different conditions.

24 **2 Methodology**

25 In this section, first, a high-level overview of the proposed framework is presented, and then
26 different stages of the framework are discussed in detail.

27 **2.1 The proposed modelling framework**

28 Figure 1 illustrates the proposed modelling framework, which is composed of three stages,
29 i.e., pre-modelling, modelling and post-modelling.



1

2 Figure 1. The proposed framework for pavement performance modelling.

3 2.2 Data preparation and feature selection

4 Pre-modelling is the most important and time-consuming stage, including data preparation
5 and feature selection. First, all the expressways need to be divided into sub-sections to ensure
6 the prediction accuracy. Two segment groups were formed, including the maintained and
7 unmaintained segments. Unmaintained segments refer to the segments that have never been
8 maintained at the time of condition data collection. Therefore, it includes two situations: 1)
9 the segment has never been maintained since it was opened to traffic; 2) the segment has been
10 maintained, but the condition data were collected before the first maintenance. Conversely,
11 maintained segments are the segments that have been maintained for one or more times and
12 the condition data were collected after the maintenance. Thus, once a segment was
13 maintained for the first time, it converted from an unmaintained segment to a maintained
14 segment. After that, data extracted from multiple databases were combined to form the panel
15 data. A method developed in the previous study (Yao *et al.*, 2019), which adopted the longest
16 increasing (for non-decreasing indexes) or decreasing (for non-increasing indexes)
17 subsequence method, was utilized to remove the noise and inconsistent data in the dataset.
18 Optimal features were selected by integrating the consideration of domain knowledge and

1 model efficiency. More specifically, an initial set of input features were chosen as candidates
2 for use by the models based on literature review, engineering judgement and available data.
3 Later, a feature selection algorithm named BorutaShap (Keany, 2020) was applied on the
4 initial feature set to automatically determine the optimal feature set for each model.

5 BorutaShap combines the Boruta feature selection algorithm with the SHapley
6 Additive exPlanations (SHAP) technique, which was proposed and written as a python
7 package by Keany (2020). It can be considered as an improvement of the original Boruta
8 algorithm. Boruta is a wrapper method built around the Random Forest model (Kursa and
9 Rudnicki, 2010). It is an all-relevant feature selection method, aiming to capture all the
10 important features in the dataset that is usable for prediction. The importance of a feature is
11 measured by the permutation importance, which is defined as the loss in accuracy of the
12 model by randomly shuffling a feature in the dataset (Kursa and Rudnicki, 2010). However,
13 this importance metric causes the algorithm computationally expensive and is not considered
14 to be a reliable measure of global feature importance, which led to the replacement of
15 permutation importance with SHAP importance in the BorutaShap algorithm (Keany, 2020).
16 Since SHAP method is also used for model interpretation, more details about it will be
17 introduced later. Besides, in this study, the BorutaShap algorithm was implemented in Python
18 using the BorutaShap package developed by Keany (2020).

19 **2.3 Model development**

20 In the modelling stage, Bayesian neural network (BNN) (Gal *et al.*, 2017), which is powerful
21 in avoiding over-fitting and estimating uncertainties, was employed to develop two pavement
22 performance models to model the performance deterioration of the maintained and
23 unmaintained segments, respectively.

24 Bayesian neural network differs from the standard deep learning in that all its
25 parameters are represented by probability distributions rather than having fixed values
26 (Blundell *et al.*, 2015). This makes BNN powerful in avoiding over-fitting problems and
27 quantifying the uncertainties in predictions. For a training data set $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$, $\mathbf{Y} =$
28 $\{y_1, y_2, \dots, y_N\}$, letting $p(\mathbf{W})$, $y = f^{\mathbf{W}}(x)$ and $p(\mathbf{Y}|\mathbf{X}, \mathbf{W})$ be the prior distribution of the
29 network parameters, the random output and the model likelihood of the BNN model, the
30 posterior over the parameters can be computed using Bayesian inference (Ryu *et al.*, 2019):

$$p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})}{p(\mathbf{Y}|\mathbf{X})} \quad (1)$$

1 The predictive distribution for a new input x^* then can be obtained as (Ryu *et al.*, 2019):

$$p(y^*|x^*, \mathbf{X}, \mathbf{Y}) = \int p(y^*|x^*, \mathbf{W}) p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) d\mathbf{W} \quad (2)$$

2 However, the posterior $p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$ cannot be solved analytically. Instead, it is fitted with an
 3 approximating variational distribution $q_\theta(\mathbf{W})$, parameterised by θ . Kullback–Leibler (KL)
 4 divergence can be used to measure the similarity between the two distributions (Gal and
 5 Ghahramani, 2015):

$$KL(q_\theta(\mathbf{W})||p(\mathbf{W}|\mathbf{X}, \mathbf{Y})) = \int q_\theta(\mathbf{W}) \log \frac{q_\theta(\mathbf{W})}{p(\mathbf{W}|\mathbf{X}, \mathbf{Y})} d\mathbf{W} \quad (3)$$

6 By replacing the intractable posterior in Eq. (3) with Eq. (1), minimizing the KL divergence
 7 becomes equivalent to maximising the log evidence lower bound (Gal and Ghahramani,
 8 2015):

$$L_{V1} := \int q_\theta(\mathbf{W}) \log p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) d\mathbf{W} - KL(q_\theta(\mathbf{W})||p(\mathbf{W})) \quad (4)$$

9 The posterior distribution is usually approximated using factorised Gaussian distributions.
 10 However, this approximation does not perform well in practice due to the large number of
 11 extra parameters that need to be learnt (Ryu *et al.*, 2019). In contrast, Monte Carlo (MC)
 12 dropout is another variational approximation method which applies dropout at both training
 13 and test time (Gal and Ghahramani, 2016; Gal *et al.*, 2017). For regression, sampling T times
 14 from the approximate posterior $q_\theta(\mathbf{W})$ leads to the estimation of the predictive mean:

$$E(y) = \frac{1}{T} \sum_{t=1}^T f^{w_t}(x) \quad (5)$$

15 , and the predictive variance (Ryu *et al.*, 2019):

$$Var(y) \approx \sigma^2 + \frac{1}{T} \sum_{t=1}^T f^{w_t}(x)^T f^{w_t}(x) - E(y)^T E(y) \quad (6)$$

1 where σ denotes the parameter of Gaussian likelihood: $p(y|f^{\mathbf{W}}(x)) = N(y; f^{\mathbf{W}}(x), \sigma)$. σ^2 is
 2 termed aleatoric uncertainty, which captures the amount of noise inherent in the outputs. It is
 3 homoscedastic for every data point but can be different or heteroscedastic for different data
 4 points. It is worth noting that it is very important to assume heteroscedastic uncertainty
 5 because the noise level in real-world data is usually unbalanced. In this case, the model output
 6 is expressed as $[y, \sigma^2] = f^{\mathbf{W}}(x)$. Both predictive mean and predictive variance can be
 7 obtained through a single network (Kendall *et al.*, 2017). Finally, the predictive uncertainty is
 8 given by (Ryu *et al.*, 2019):

$$\begin{aligned}
 \text{Var}(y) \approx & \underbrace{\frac{1}{T} \sum_{t=1}^T y_t^2 - \left(\frac{1}{T} \sum_{t=1}^T y_t \right)^2}_{\text{epistemic}} + \underbrace{\frac{1}{T} \sum_{t=1}^T \sigma_t^2}_{\text{aleatoric}} \quad (7)
 \end{aligned}$$

9 where $y_t, \sigma_t = f^{w_t}(x)$ and w_t is randomly sampled from $q_{\theta}(\mathbf{W})$. Eq. (7) can be split into
 10 epistemic uncertainty and aleatoric uncertainty as labelled. Aleatoric uncertainty is data-
 11 driven and cannot be reduced by collecting more data while epistemic uncertainty is model-
 12 driven and can be explained given enough data (Kendall *et al.*, 2017). The models were
 13 trained and tested in Python 3.8.3 using the codes adapted from
 14 <https://github.com/anassinator/bnn>.

15 2.4 Model interpretation

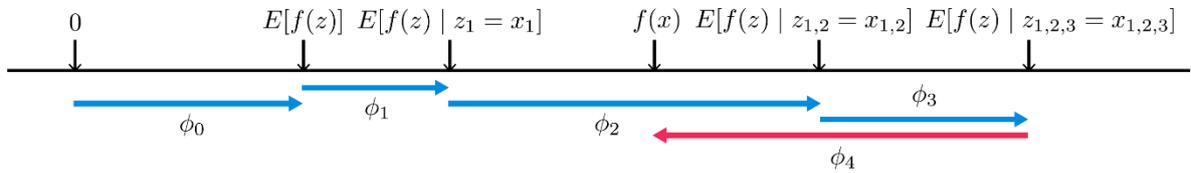
16 Post-modelling, generally referring to model interpretation, is a stage that was often
 17 overlooked but actually indispensable in pavement performance modelling. In the developed
 18 framework, the SHapley Additive exPlanations (SHAP) approach (Lundberg and Lee, 2017)
 19 was applied to the established BNN models to investigate the feature importance and feature
 20 effect.

21 SHAP is an approach based on coalitional game theory to explain the output of any
 22 machine learning models recently proposed by Lundberg and Lee (2017). Shapley values
 23 take each feature value of the data instance as the player and the prediction as the “payout”,
 24 and then consider how to distribute the “payout” fairly among the different features (Molnar
 25 2018). Thus, the Shapley value can be regarded as the average marginal contribution of a
 26 feature value across all possible coalitions. SHAP goes further by connecting different

1 explanation methods with Shapley values, representing the Shapley value explanation as an
 2 additive feature attribution method, i.e., a linear model. Let $f(x)$ and $g(x')$ be the prediction
 3 model and explanation model, respectively, such that $x = h_x(x')$ is the mapping function that
 4 maps the simplified inputs x' to the original input x , then the explanation model $g(x')$ for the
 5 original model $f(x)$ can be expressed as:

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^M \phi_i x_i' \quad (8)$$

6 where M is the number of simplified inputs and ϕ_i is the SHAP value, which indicates the
 7 effect of each feature. The effects of all the features are added up to approximate the output
 8 of the original model. Figure 2 illustrates how the prediction develops from the base value
 9 $E[F(z)]$, which is predicted without knowing any feature values, to the current output $f(x)$
 10 based on the SHAP values of each feature. In other words, the SHAP value of each feature is
 11 just like a force to either increase or decrease the prediction. More details about SHAP can be
 12 found in (Lundberg and Lee, 2017).



13

14 Figure 2. SHAP values (Lundberg and Lee, 2017).

15 According to the different approximation methods of SHAP values, the SHAP
 16 approach can be further classified into Kernel SHAP, Tree SHAP and Deep SHAP (Lundberg
 17 *et al.* 2020; Shrikumar *et al.*, 2017). Kernel SHAP estimates SHAP values for any model
 18 using a specially-weighted local linear regression. Tree SHAP is mainly for tree-based
 19 models, such as decision trees, random forest (RF), etc. It is also embedded in the Brutashap
 20 algorithm as Brutashap is built around the RF model. Deep SHAP, on the other hand, can
 21 generate efficient approximation of SHAP values in deep learning models. Hence, Deep
 22 SHAP is used to interpret the BNN model output in this study.

23 This begins with the computation of the SHAP value, which could be regarded as the
 24 marginal contribution of every feature for every sample. The feature importance was then
 25 measured by averaging the SHAP values over all samples. By plotting the SHAP value of the

1 feature versus the value of the feature for all examples in the dataset and colouring the points
2 by another feature, the interaction effect between features could be visualized. All these
3 analyses were carried out in Python using the “shap” package.

4 **3 Case Study**

5 While the framework proposed in this study is applicable to various pavement performance
6 indicators, the Transverse Cracks Evaluation Index (TCEI) was selected to conduct the case
7 study, as transverse cracking is one of the major distress modes for semi-rigid base asphalt
8 pavement, which is the most popular type of pavement in China. The TCEI was calculated by
9 normalizing the ratio of transverse crack spacing (TCS) to transverse crack width ratio (TWR)
10 to a range of 0 to 100, with 100 representing the best condition (i.e., no transverse cracks). It
11 was proposed by Zhou *et al.* (2010) and has been widely used in Jiangsu and Anhui province,
12 China (Han *et al.*, 2020; Yao *et al.*, 2019).

13 The data for this case study were from the pavement management system of Jiangsu
14 Province, China, covering a total of 49 expressways. All the expressways were first divided
15 into sub-sections according to their directions, lanes, locations, structure sections, and traffic
16 sections. Then, the sub-sections longer than 1.5 km were further divided at intervals of 1 km.
17 The lengths of pavement segments after the completion of roadway segmentation were
18 between 0.1 km to 1.5 km. The data with only basic road section information were then
19 expanded to create multi-feature, multi-observation panel data. 35,599 road segments with a
20 total length of 14,162 lane-kilometres and 258,988 observations were acquired. Among them,
21 195,921 observations were collected from the unmaintained segments while 63,067
22 observations were collected from the maintained segments.

23 In this study, the performance model was formulated as a recursive nonlinear model,
24 which was considered to be more beneficial to pavement management since the condition
25 data were usually available on a regular basis (Prozzi and Madanat, 2003). The general form
26 of the model can be expressed as follows:

$$y_{t+1} = f(y_t, \text{structure}, \text{material}, \text{traffic}, \text{climate}, \text{maintenance} \dots) \quad (9)$$

27 where y_{t+1} and y_t are the pavement conditions at time t+1 and t, respectively; and structure,
28 material, traffic, climate and maintenance denote the different influential factors. The model
29 would predict the TCEI of the following year using the TCEI of the previous year as well as
30 other features. However, in some cases, if the pavement inspection was not carried out every

1 year or some of the inspection data were missing or considered abnormal, it was also possible
 2 to predict the TCEI in a few years. Therefore, time increment was also among the considered
 3 features.

4 Preliminary selection of the features was based on literature review (Yao *et al.*, 2019;
 5 Karlaftis and Badr, 2015; Onayev and Swei, 2021) and data availability. As Table 1 shows,
 6 38 features covering different characteristics of the pavement were selected. Except for
 7 ESAL, all other features related to traffic load and climate condition were expressed in the
 8 form of mean values or ratios during the period from time t to time $t+1$. Features in the
 9 category of maintenance history are only included in the analysis of maintained segments. *No.*
 10 *of lanes&lane* is a combination of two features: number of lanes in both directions (4, 6, or 8)
 11 and the lane (the 1st, 2nd, 3rd, or 4th lane outward from the central divider, if applicable). This
 12 feature was used as a proxy of lane distribution factor, depicting the traffic distribution in
 13 different lanes. *R_{below 0 °C}* and *R_{over 35 °C}* represent the ratio of days with daily
 14 minimum temperature below 0 °C and daily maximum temperature over 35 °C, respectively.
 15 *R_{greater_or_equal_10 mm}* is the ratio of days with 24-hour precipitation greater than or
 16 equal to 10 mm. ESAL, MESAL and MADT denote equivalent single-axle loads, monthly
 17 equivalent single-axle loads, and monthly average daily traffic, respectively. In addition, the
 18 different ages of different asphalt layers are used to reflect the rehabilitation history of the
 19 maintained segments.

20 Table 1 The initial feature set pending for selection.

Feature type	Feature name
Basic information	Current TCEI
	Road age
	Pavement or bridge
	Time increment
	No. of lanes&lane
Pavement structure and material	Styrene-Butadiene-Styrene (SBS) modified asphalt layer thickness
	Asphalt layer materials
	Asphalt layer thickness
	Base layer material
	Base layer thickness
Traffic load	Number of single-axle loads
	Single-axle overload rate
	Number of tandem-axle loads
	Tandem-axle overload rate
	Number of tridem-axle loads
	Tridem-axle overload rate
	ESAL
	ESAL increment
Mean MESAL	

	Mean MADT
	Ratio of passenger cars to trucks
Climate condition	Mean temperature
	Mean daily precipitation
	Mean relative humidity
	Mean wind speed
	Ratio of rainy days
	R_below 0 °C
	R_over 35 °C
	Extreme temperature range
	Maximum 5 d mean temperature
	Maximum 5 d precipitation total
	R_greater or equal 10 mm
	Pre-treatment TCEI
Maintenance history	Treatment age
	Maintenance and Rehabilitation (M&R) treatment
	Upper asphalt layer age
	Middle asphalt layer age
	Lower asphalt layer age

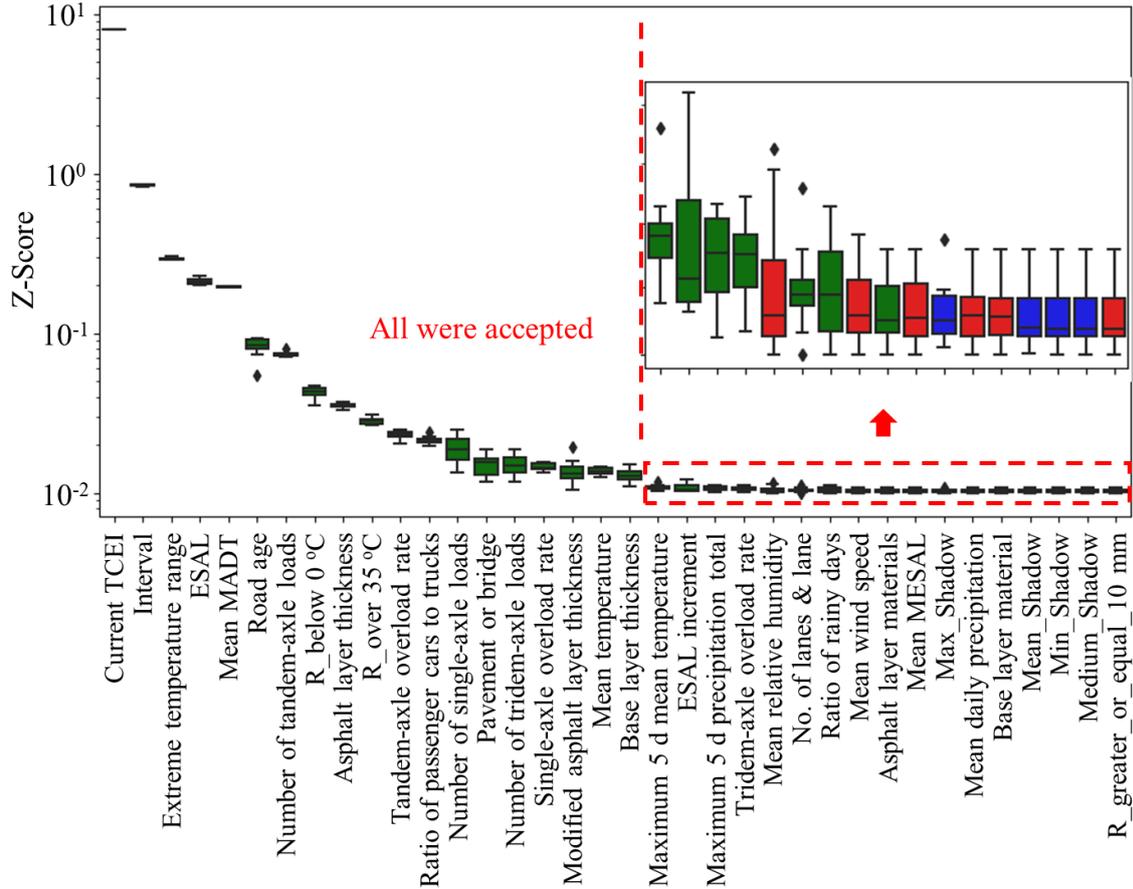
1

2 **4 Results and Discussion**

3 **4.1 Selected features**

4 **4.1.1 Unmaintained segments**

5 Figure 3 shows the feature selection results for unmaintained segments. Each feature is
6 represented by a boxplot. Green and red boxes indicate the features that have been accepted
7 and rejected, respectively, while blue boxes represent the shadow features. The y axis is a
8 metric of feature importance in logarithmic scale. A total of 26 features were selected for
9 building the BNN model. Among them, the current TCEI is the most dominant. The
10 descriptive statistics of the optimal features are also shown in Table 2. Better performance
11 and higher reliability of the model are expected for the cases within the ranges given in Table
12 2.



1

2 Figure 3. Results of feature selection for unmaintained segments.

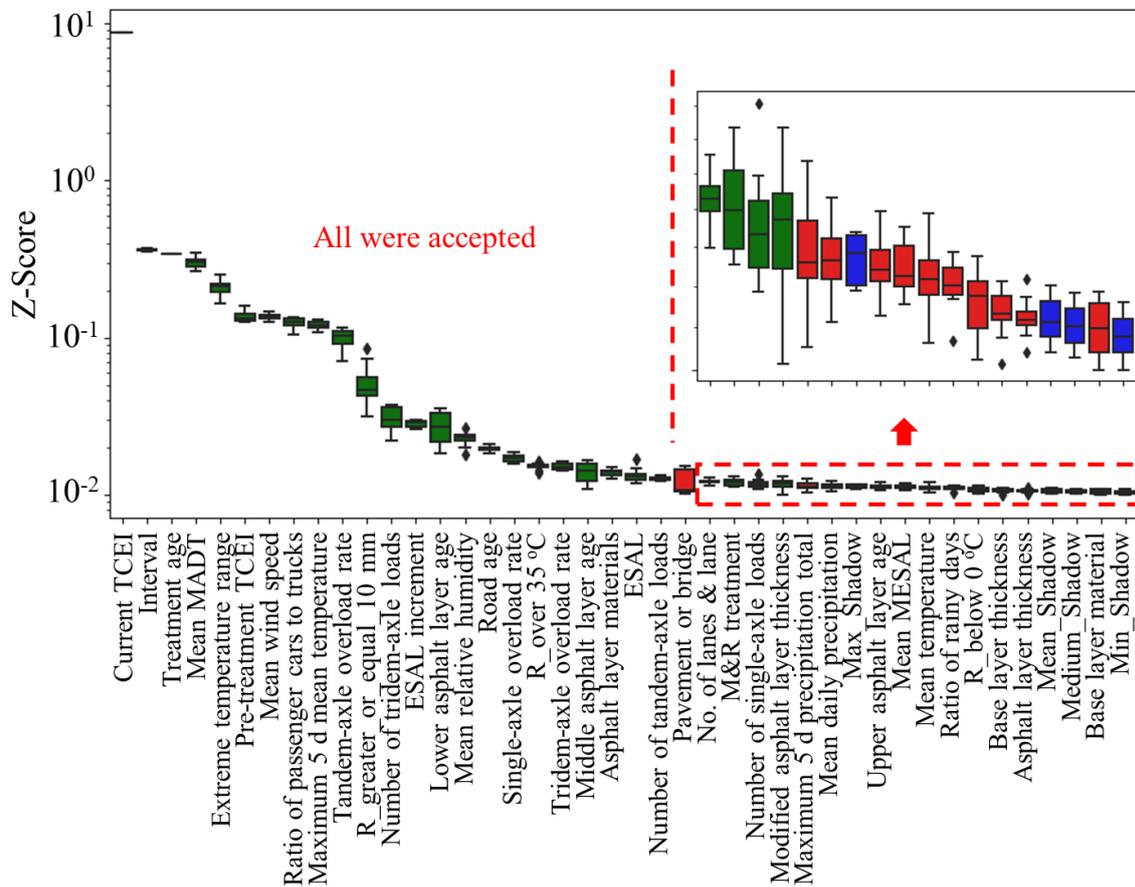
3 Table 2 Descriptive statistics of the selected features for unmaintained segments.

Variables	Mean	Std	Min.	Max.	Median	Unit
Current TCEI	93.8	13.3	15.3	100.0	100.0	/
Time increment	1.2	0.9	0.1	20.0	1.0	year
Extreme temperature range	46.0	3.8	18.4	54.2	45.9	°C
ESAL	777.1	904.4	0.0	10470.0	482.5	× 10 ⁴
Mean MADT	10737.3	19225.6	0.0	256628.8	6119.2	/
Road age	5.7	3.5	0.0	20.2	5.4	year
Number of tandem-axle loads	29055.4	26869.9	32.2	422194.0	20271.2	/
R_below 0 °C	18.9	8.3	1.0	90.3	18.2	%
Asphalt layer thickness	18.1	1.1	16.0	21.0	18.0	cm
R_over 35 °C	1.9	2.0	0.0	15.8	1.2	%
Tandem-axle overload rate	38.7	17.5	5.7	94.0	35.6	%
Ratio of passenger cars to trucks	1.9	1.0	0.2	15.4	1.8	/
Number of single-axle loads	64330.2	64822.8	247.2	1212853.2	45004.0	/
Pavement or bridge	/	/	/	/	/	/
Number of tridem-axle loads	31964.9	31579.6	20.2	574032.4	22403.8	/
Single-axle overload rate	40.4	16.3	4.9	80.8	38.0	%
Modified asphalt layer thickness	9.2	4.0	0.0	20.0	10.0	cm
Mean temperature	14.5	2.6	-1.0	20.9	14.8	°C
Base layer thickness	37.6	1.6	0.0	46.0	38.0	cm
Maximum 5 d mean	30.0	3.4	3.0	35.0	30.6	°C

temperature						
ESAL increment	161.2	229.6	0.0	5757.7	91.3	$\times 10^4$
Maximum 5 d precipitation total	191.2	89.1	3.5	455.7	183.5	mm
Tridem-axle overload rate	59.6	15.9	9.6	95.9	57.7	%
No. of lanes & lane	/	/	/	/	/	/
Ratio of rainy days	27.1	5.5	6.5	46.7	26.0	%
Asphalt layer materials	/	/	/	/	/	/

1 4.1.2 Maintained segments

2 For the maintained segments, a subset of 27 features were selected for further analysis, as
 3 illustrated in Figure 4 The current TCEI is still the most important. The descriptive statistics
 4 of the selected features are shown in Table 3.



5
 6 Figure 4. Results of feature selection for maintained segments.

7 Table 3 Descriptive statistics of the selected features for maintained segments.

Variables	Mean	Std	Min.	Max.	Median	Unit
Current TCEI	92.8	14.4	22.1	100.0	100.0	/
Time increment	1.1	0.6	0.1	12.4	1.0	year
Treatment age	3.5	2.9	0.0	14.4	2.8	year
Mean MADT	14135.6	17442.7	0.0	716098.0	10644.8	/

Extreme temperature range	46.3	3.9	17.3	53.3	47.0	°C
Pre-treatment TCEI	85.5	20.3	6.9	100.0	100.0	/
Mean wind speed	2.3	0.3	1.0	3.5	2.3	m/s
Ratio of passenger cars to trucks	1.8	1.0	0.2	15.1	1.8	/
Maximum 5 d mean temperature	30.7	3.1	3.0	35.0	31.1	°C
Tandem-axle overload rate	32.7	15.3	7.5	90.4	29.1	%
R_greater_or_equal_10 mm	8.0	2.4	0.0	26.2	8.0	%
Number of tridem-axle loads	64759.6	46787.7	336.0	574032.4	56278.9	/
ESAL increment	221.1	253.2	0.2	5257.4	155.3	$\times 10^4$
Lower asphalt layer age	7.9	5.0	0.0	20.7	7.4	year
Mean relative humidity	72.0	4.0	50.0	86.1	73.0	%
Road age	11.6	3.8	0.7	20.7	11.9	year
Single-axle overload rate	39.2	15.8	8.8	80.0	36.6	%
R_over 35 °C	2.3	2.2	0.0	29.6	1.6	%
Tridem-axle overload rate	61.5	15.5	15.1	94.8	59.9	%
Middle asphalt layer age	9.4	4.5	0.0	20.7	10.0	year
Asphalt layer materials	/	/	/	/	/	/
ESAL	3029.5	2615.4	0.0	11408.7	2135.8	$\times 10^4$
Number of tandem-axle loads	46322.4	31031.4	361.4	422194.0	44552.1	/
No. of lanes & lane	/	/	/	/	/	/
M&R Treatment	/	/	/	/	/	/
Number of single-axle loads	104864.7	80854.5	622.6	1212853.2	93972.3	/
Modified asphalt layer thickness	6.5	4.6	0.0	18.0	4.0	cm

1 4.2 Model training

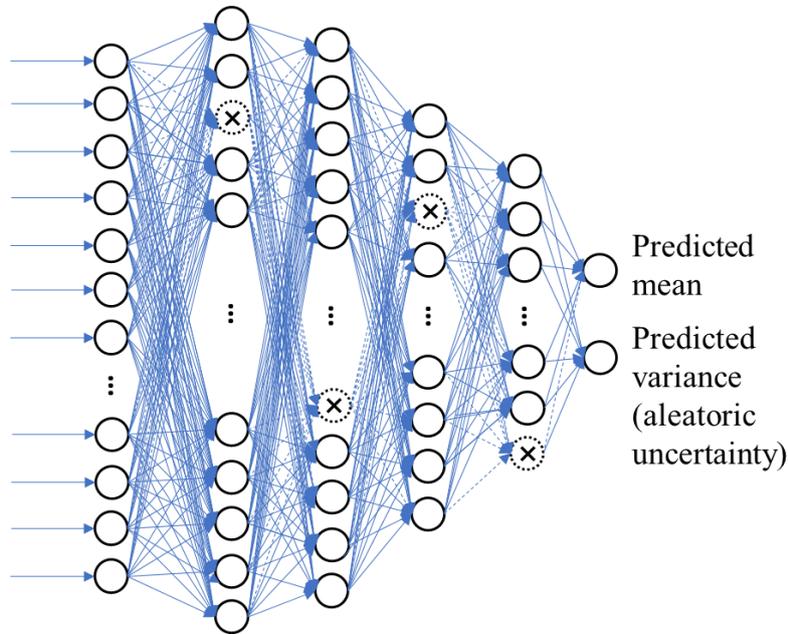
2 To develop the BNN model, the data was first normalized through one-hot coding (for
3 categorical variables) and standardization (for numerical variables). The entire data set was
4 then split in a ratio of 75:25, i.e., 75% of the data were used for training the model while 25%
5 were used for testing the model that was built out of it. Hyperparameters, such as the number
6 of hidden layers and neurons, optimizer, learning rate, batch size and number of epochs were
7 tuned manually through trial and error. Table 4 shows the optimum hyperparameters that
8 were acquired after careful tuning and utilized to build the BNN models. $n_particles$ is the
9 number of particles used to estimate the epistemic uncertainty, i.e., epistemic uncertainty was
10 estimated by running the model $n_particles$ times for each input sample. Both models have
11 one input layer, four hidden layers and one output layer. The configuration of the optimal
12 model is 26-128-128-64-16-2 for unmaintained segments and 27-256-256-64-16-2 for
13 maintained segments, with the numbers denoting the neuron numbers in each layer. Figure 5
14 shows the schematic diagram of the model structures for unmaintained and maintained
15 segments. The circles and arrows indicate the neurons and connections, respectively. A
16 dotted circle with a cross in the centre stands for those dropped units and the dashed arrows

1 are invalid neuronal connections. As dropout was applied at both training and test time,
 2 certain percent of units would be randomly dropped in each run, allowing the estimation of
 3 epistemic uncertainty. For more information about BNN model training, the reader can refer
 4 to the original article (Gal et al., 2017).

5 Table 4 Model hyperparameters.

	Maintained segments	Unmaintained segments
Model structure	27-256-256-64-16-2	26-128-128-64-16-2
Optimizer	Adam	Adam
Learning rate	0.0001	0.0001
Batch size	128	128
Epochs	500	500
n_particles	100	100
Activation function	Rectified Linear Unit (ReLU)	ReLU

6



7

8 Figure 5. A schematic diagram of the BNN model structure.

9 After the model was trained, both epistemic and aleatoric uncertainty could be
 10 quantified. Aleatoric uncertainty arise from data noise, which is input-dependent and could be
 11 directly output from the model. Epistemic uncertainty is related to the uncertainty in the
 12 model parameters. It was estimated by running the model for multiple times.

13

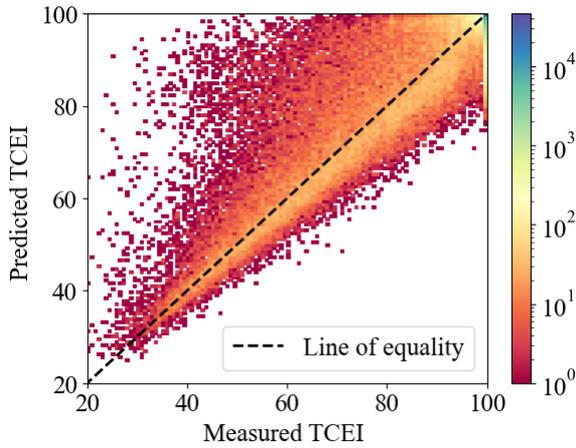
1 4.3 Model performance

2 The performances of the two models were evaluated using the metrics in Table 5. R-square is
 3 the statistical parameter that represents the proportion of the variance for a dependent
 4 variable that is explained by the independent variables. Mean Absolute Error (MAE)
 5 measures the average magnitude of the errors in the predictions, without considering their
 6 direction. Root Mean Squared Error (RMSE) is the square root of the average squared errors,
 7 which assigns a much larger weight to larger errors compared with MAE. From Table 5, it
 8 can be observed that the model of unmaintained segments outperforms the model of
 9 maintained segments in terms of these three metrics for both the training and testing dataset.
 10 This is expectable since the implementation of different maintenance activities may make the
 11 deterioration of pavement performance more complicated and unpredictable. Meanwhile, the
 12 much smaller sample size of the model of maintained segments may also has some effect on
 13 its relatively lower accuracy. In addition, Figure 6 shows the two-dimensional histograms of
 14 the measured TCEI verse the predicted TCEI with the colour of the points representing the
 15 density of the points at a given location. It implies that although most of the points are
 16 densely around the line of equality, there are still some points that are not well predicted
 17 which also explains the necessity of uncertainty analysis. But overall, the performances of
 18 these two models are comparable with recently developed ML-based models (Kırbaş and
 19 Kardeş, 2016; Madeh Piryonesi and El-Diraby, 2021; Yao *et al.*, 2019). Moreover, the
 20 much larger dataset used in this study allow the model to handle more complex real-world
 21 problems. The ensemble learning in BNN also ensures the model to achieve higher
 22 generalization performance (Barber and Bishop, 1998; Wittek, 2014).

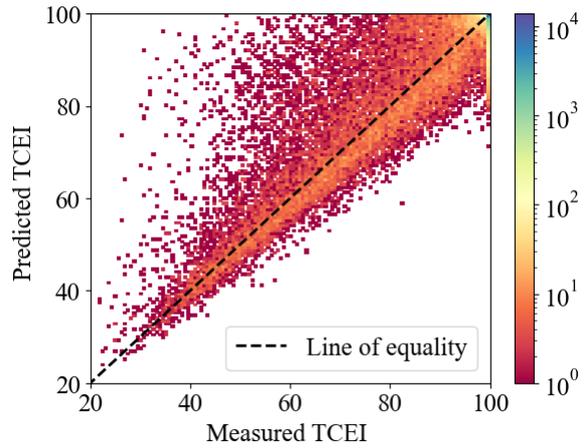
23 Table 5 Model evaluation results.

Dataset	Evaluation metrics	Unmaintained segments	Maintained segments
Training	Sample size	137649	40759
	R-square	0.86	0.79
	MAE	3.08	5.11
	RMSE	6.21	8.50
Testing	Sample size	45884	13587
	R-square	0.85	0.77
	MAE	3.13	5.37
	RMSE	6.26	8.92

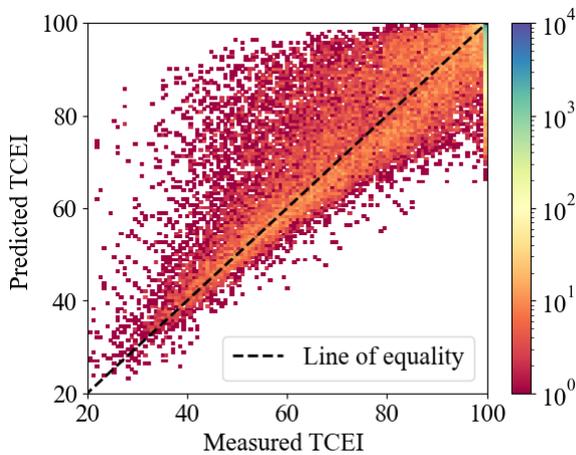
24



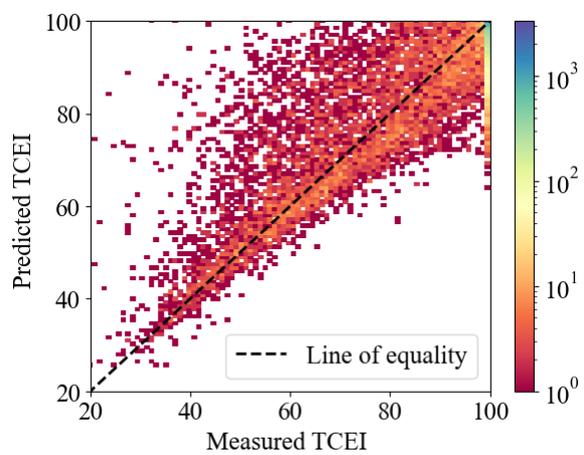
(a) Training data of unmaintained segments



(b) Testing data of unmaintained segments



(c) Training data of maintained segments



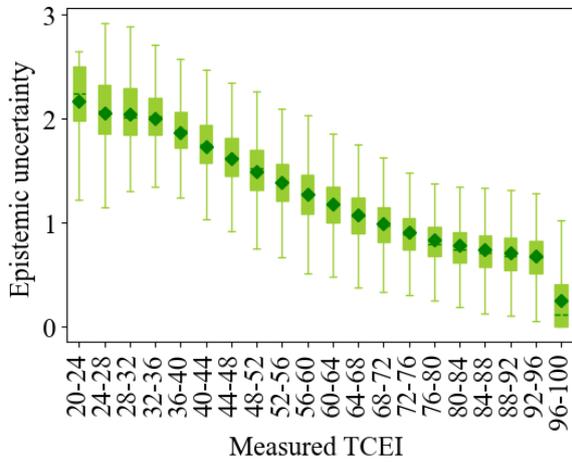
(d) Testing data of maintained segments

1 Figure 6. Measured against predicted TCEI values.

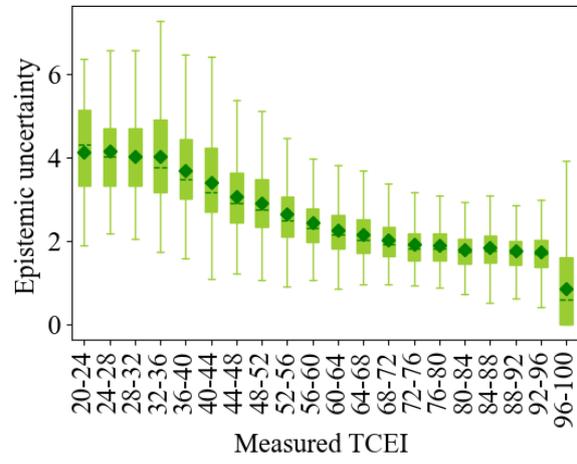
2 **4.4 Uncertainty analysis**

3 Figure 7 shows the distributions of the three types of uncertainty at different TCEI levels
 4 which were quantified by the BNN models. In general, the results of maintained segments
 5 have greater uncertainty compared with their unmaintained counterparts, but the overall
 6 trends on the boxplots are similar. Segments with lower TCEI values show larger epistemic
 7 uncertainty. One possible reason is that agencies do not allow very dense transverse cracks on
 8 the highway pavement, resulting in fewer data samples at lower TCEI values, which in turn
 9 leads to greater epistemic uncertainty. Regarding the aleatoric uncertainty, Figure 7(c) and (d)
 10 clearly show that the magnitude of data-driven uncertainty varied greatly with different TCEI

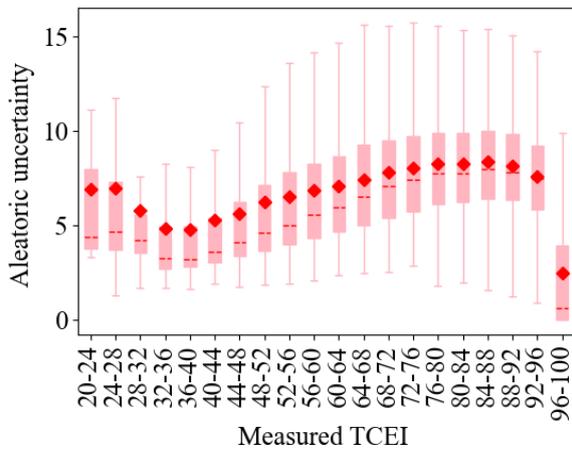
1 values. This reveals that the inherent noise in the measured TCEI was heterogenous. More
 2 attention need be paid to the inspection data with high aleatoric uncertainty. Moreover, since
 3 aleatoric uncertainties are more dominant than epistemic uncertainties, the distributions of
 4 total uncertainties are very similar to those of aleatoric uncertainties. Thus, it can be
 5 concluded that the uncertainty quantification enables to identify the uncertainty sources,
 6 evaluate the data quality, and estimate the prediction confidence. It also provides implications
 7 for reducing the uncertainty and improving the prediction. For example, the results of this
 8 study suggest that continuously improving the quality of inspection data and expanding the
 9 data to include more observations from segments with poor transverse cracking conditions
 10 are two potential approaches to enhance the reliability of prediction.



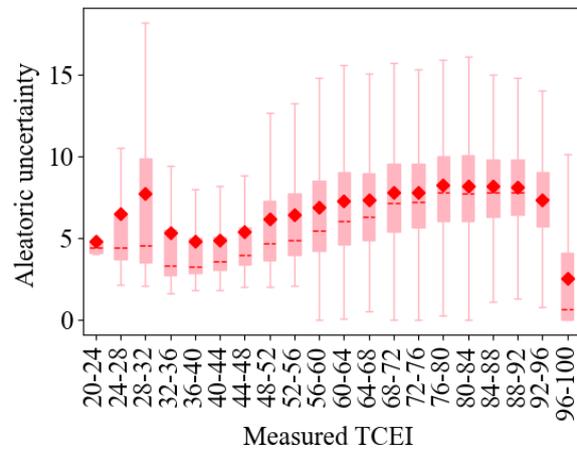
(a) Epistemic uncertainties of unmaintained segments



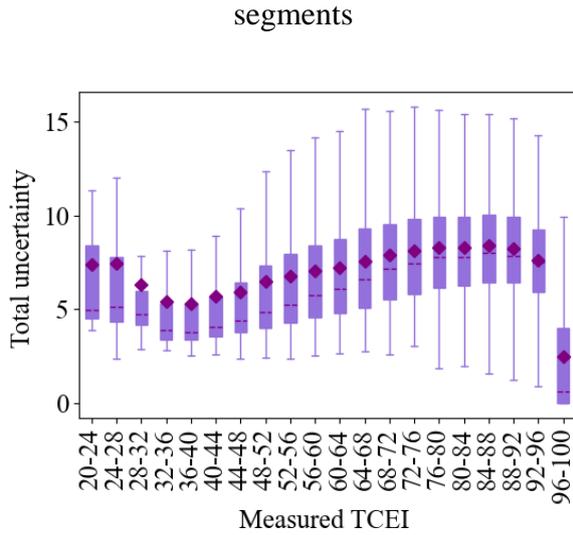
(b) Epistemic uncertainties of maintained segments



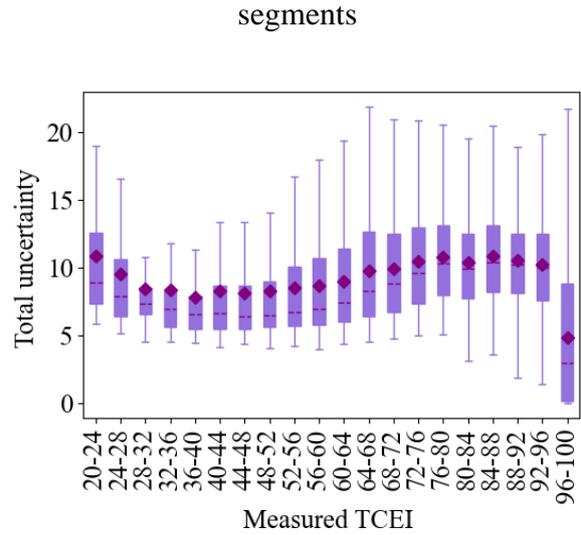
(c) Aleatoric uncertainties of unmaintained segments



(d) Aleatoric uncertainties of maintained segments



(e) Total uncertainties of unmaintained segments



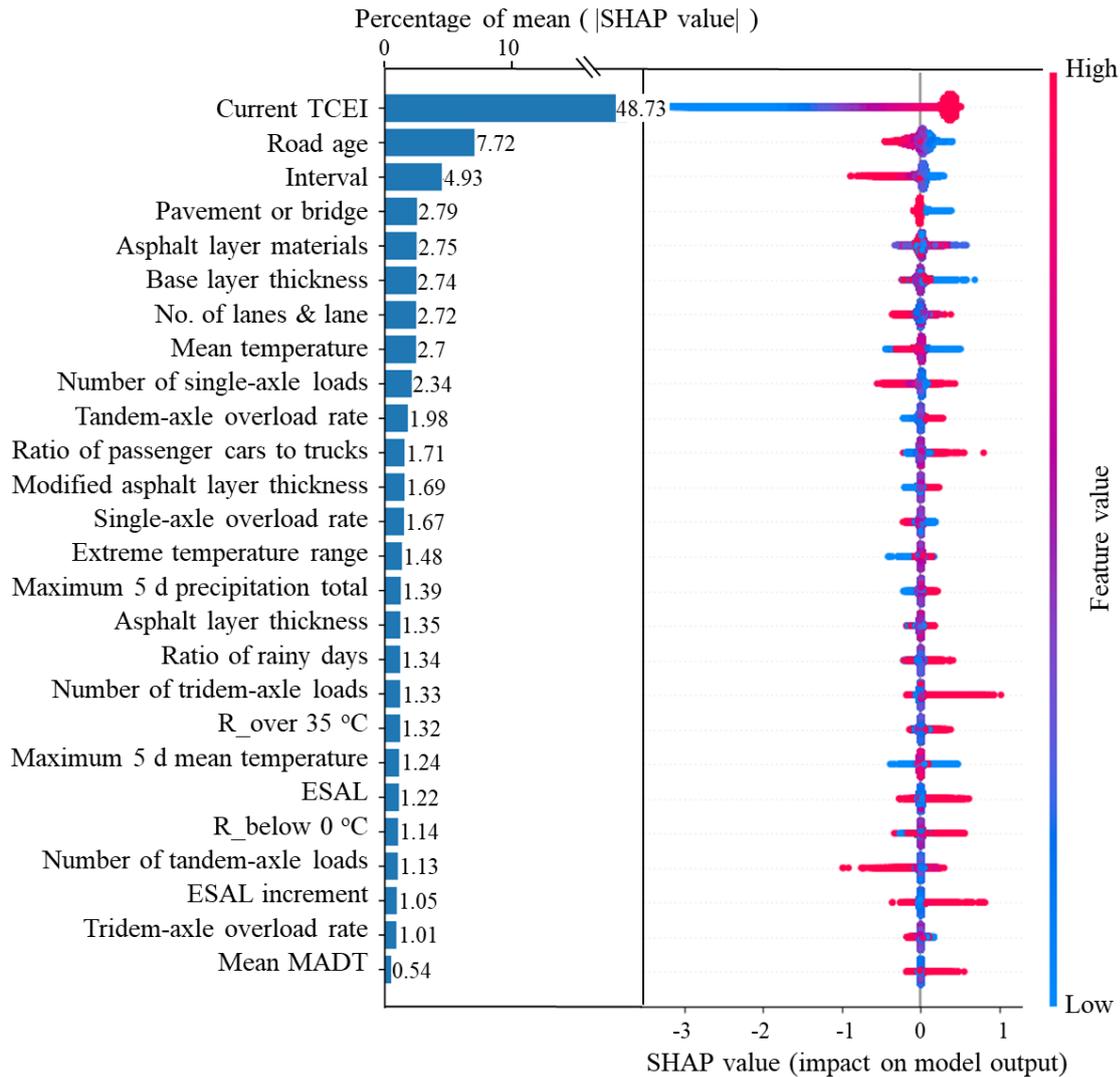
(f) Total uncertainties of maintained segments

1 Figure 7. Epistemic, aleatoric and total uncertainties against the measured TCEI values of the
 2 entire dataset.

3 4.5 Model interpretation

4 4.5.1 Unmaintained segments

5 The left side of Figure 8 shows the importance of each feature with longer bar indicating
 6 higher importance level. The current TCEI has the dominant importance among all the
 7 features making almost half of the contributions to the prediction. The right side of Figure 8
 8 indicates the feature effect. Each point corresponds to a Shapley value of a feature and an
 9 instance. The position of the points on the x-axis is determined by the Shapley value and the
 10 different colours represent the value of each feature from low to high. From this informative
 11 figure, the relationship between the value of a feature and its impact on the prediction could
 12 be preliminary inferred. For instance, a lower current TCEI value and a higher road age value
 13 would both decrease the predicted TCEI, which corresponds to more severe transverse
 14 cracking condition. However, for the features that do not meet general engineering
 15 expectations at first glance but are ranked relatively high on the importance plot, a further in-
 16 depth exploration needs to be performed to check whether it is caused by the interaction
 17 effect between features or the few abnormal instances. Meanwhile, some features of interest
 18 also deserve further examination.



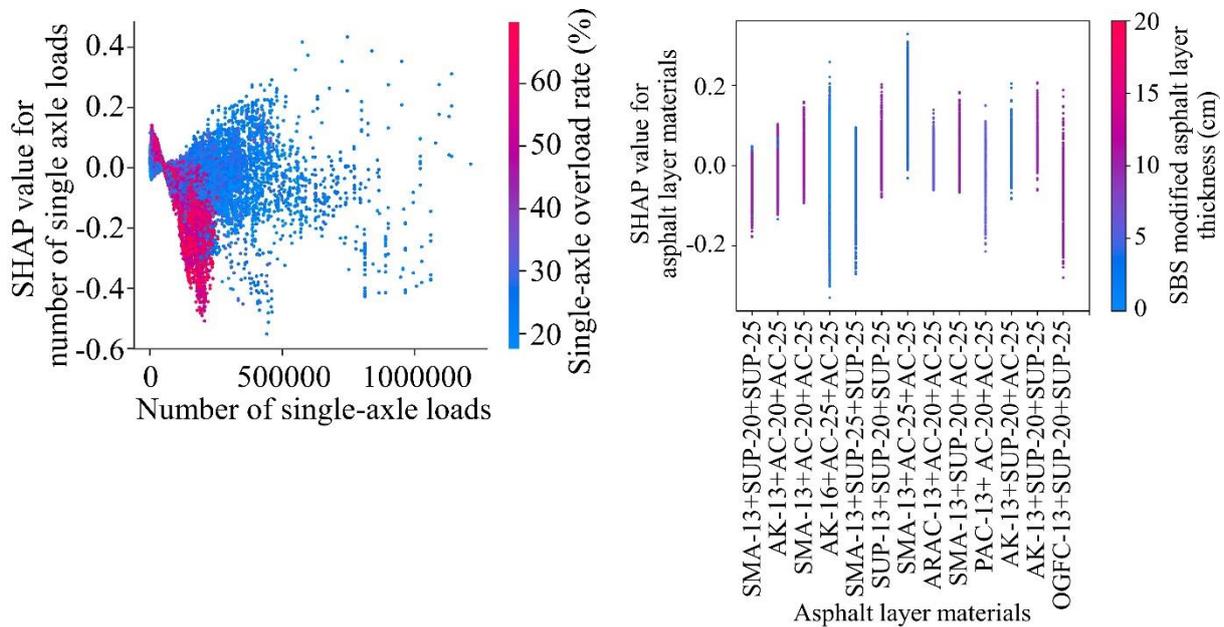
1

2 Figure 8. Results of feature importance and feature effect for unmaintained segments.

3 In order to gain an insight into those unexpected or unclear feature effects, a feature
 4 dependency analysis was conducted by plotting the variation in SHAP value with the change
 5 in the feature value. As SHAP value characterizes the marginal contribution of a feature to
 6 the change in the model output, these figures represent the change in the model output as the
 7 feature changes (Lundberg and Lee, 2017). By colouring with another feature, the interaction
 8 effect could be revealed.

9 Two features were selected as the examples. Figure 9(a) shows that the strongest
 10 interacting feature of the number of single-axle loads is its corresponding overload rate. Red
 11 points represent higher values of single-axle overload rate while blue points represent lower
 12 ones. Focusing on the red points, it can be found that under high overload rate, larger amount

1 of single-axle loads generally lead to lower predicted TCEI. However, when the overload rate
 2 is less than 30%, a larger number of single-axle loads does not necessarily decrease the
 3 prediction. Moreover, in Figure 8, asphalt layer material is represented as a dummy variable,
 4 so its effect on the model output is not clear. In this study, asphalt layer material refers to the
 5 combination of the aggregate gradation of the three asphalt layers due to no reliable material
 6 property data. SMA is a gap-graded mixture, whereas AC, AK and SUP represent three
 7 different dense-graded mixtures. ARAC, OGFC and PAC denote asphalt-rubber concrete,
 8 open-graded friction course mixture and porous asphalt concrete, respectively. The numbers,
 9 such as 13, 16, 20, and 25, signify the nominal maximum aggregate size in millimetres.
 10 Figure 9(b) illustrates the interaction of asphalt layer materials with SBS modified asphalt
 11 layer thickness. Among all the combinations with an SBS modified asphalt layer thickness of
 12 about 10 cm (purple points), the predicted TCEI corresponding to the OGFC-13 upper asphalt
 13 layer tends to be the lowest. However, when the SBS modified asphalt layer thickness is
 14 between 0 to 5 cm (blue points), the combinations of AK-16+AC-25+AC-25 and SMA-
 15 13+SUP-25+SUP-25 seem to have the largest negative contribution to the model output. All
 16 these demonstrate that looking at one single feature is usually not sufficient.



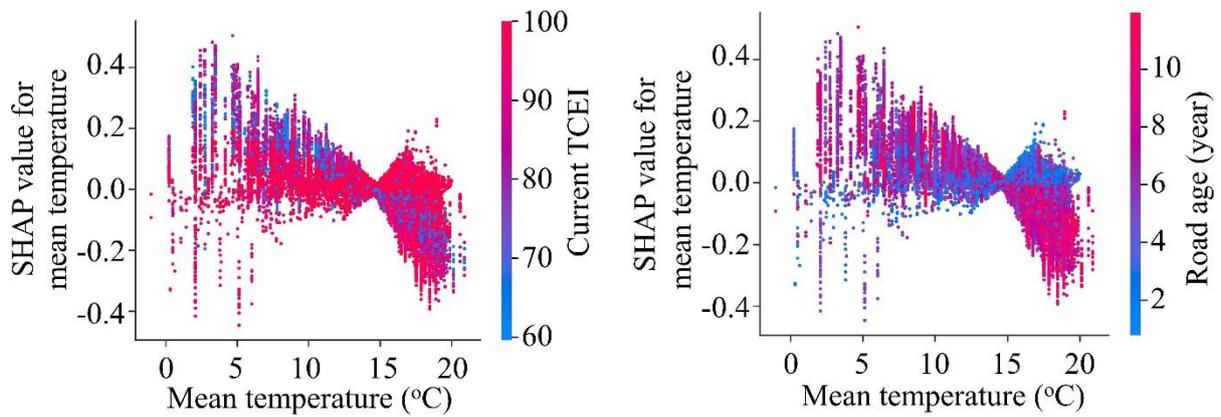
(a) Number of single-axle loads vs Single-axle overloading rate

(b) Asphalt layer materials vs SBS modified asphalt layer thickness

17 Figure 9. Visualization of feature interaction effect.

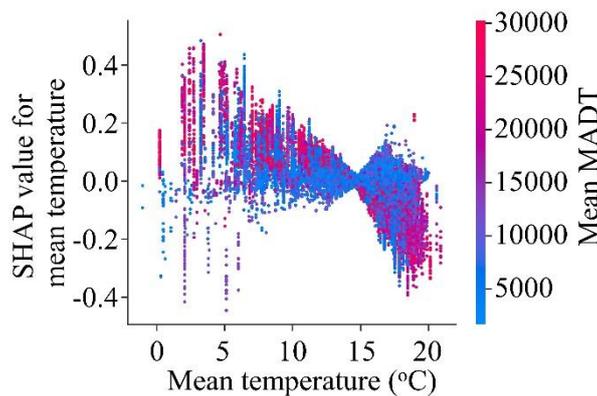
18 Furthermore, it is interesting to notice that a higher mean temperature seems to lower

1 the predicted TCEI. While digging into how mean temperature interacts with others, the
2 current TCEI, road age, and mean MADT were found to have strong interaction effect with it.
3 More specifically, the blue points in Figure 10(a) and red points in Figure 10(b) and (c)
4 suggest that a higher mean temperature has greater adverse impact on the transverse cracking
5 of asphalt pavement for road segments with lower current TCEI, larger road age, or higher
6 traffic volume. A possible reason is that exposure to a relatively high mean temperature
7 would significantly accelerate the aging of asphalt mixtures and asphalt binder hardening
8 (Braswell *et al.*, 2020). As a result, the crack resistance of asphalt pavement becomes poorer.
9 When poor current transverse cracking condition, large road age, or high traffic volume is
10 combined with high mean temperature, the aged pavement materials are more prone to
11 cracking.



(a) Mean temperature vs Current TCEI

(b) Mean temperature vs Pavement age

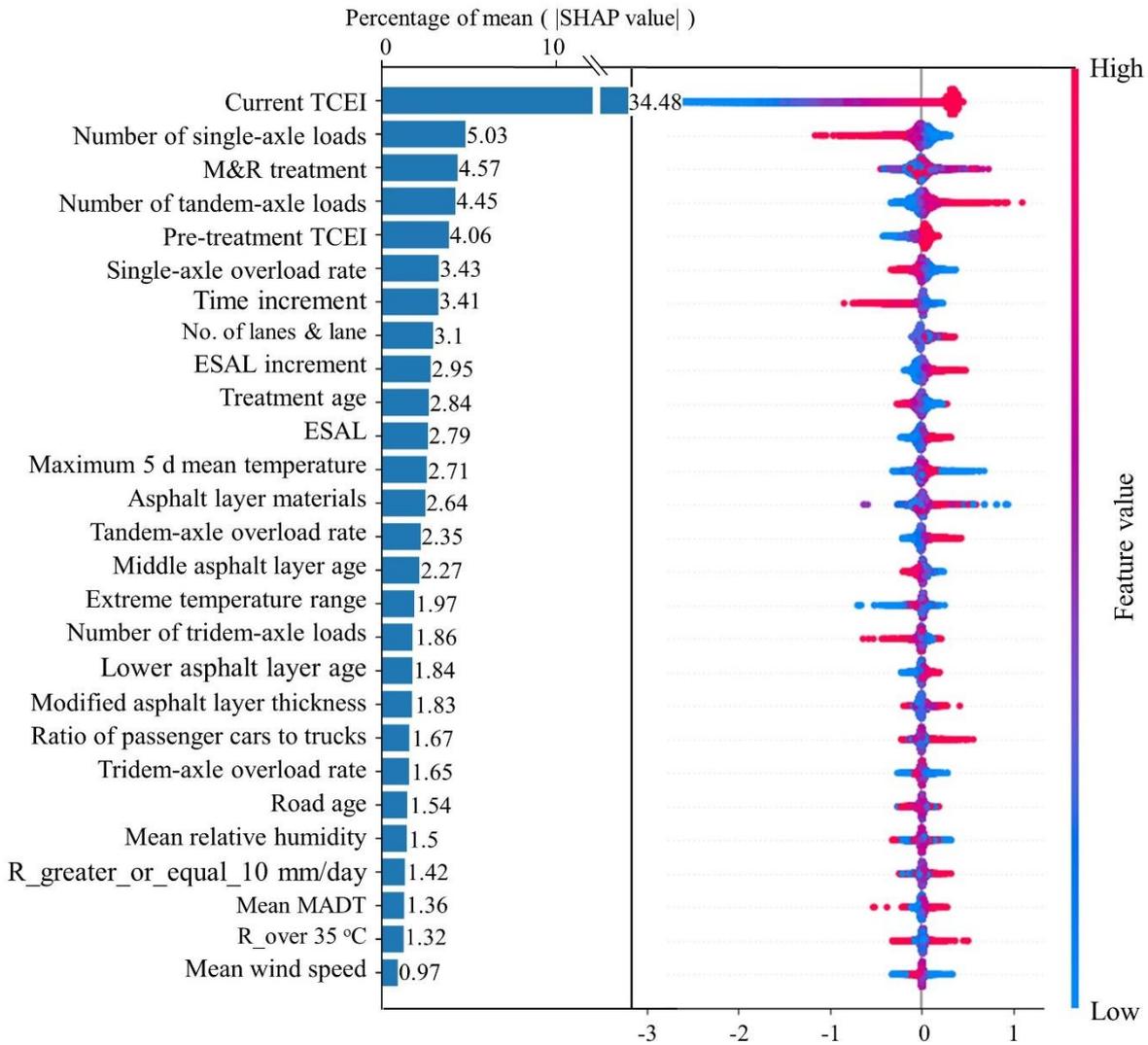


(c) Mean temperature vs Mean MADT

12 Figure 10. Visualization of interacting features of mean temperature.

1 4.5.2 Maintained segments

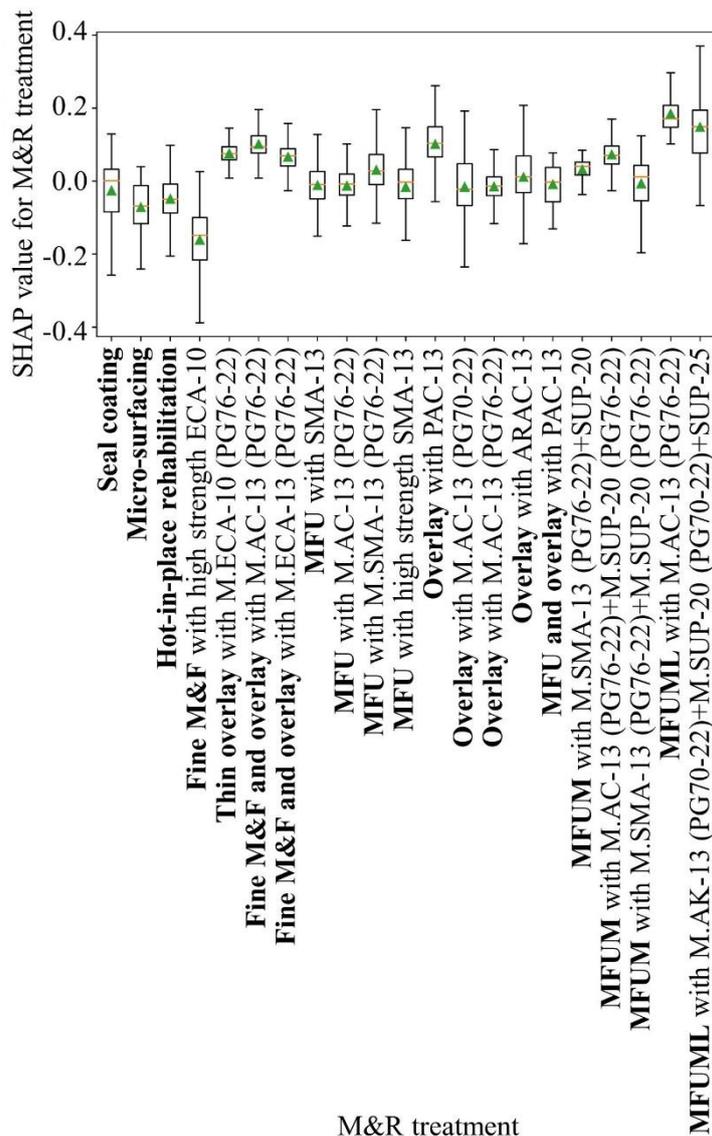
2 Figure 11 shows the feature importance and feature effect on model output for maintained
 3 segments. Current TCEI is still the most important feature. Some unique features of
 4 maintenance segments, such as M&R treatment and pre-treatment TCEI, are also of great
 5 importance.



6
 7 Figure 11. Results of feature importance and feature effect for maintained segments.

8 M&R treatment is the third important feature of the maintained model. Figure 12
 9 shows the distribution of the contribution of each M&R treatment to the model output,
 10 covering about 78% of the total M&R mileage in Jiangsu Province. Each M&R treatment is
 11 represented by the M&R type and the corresponding M&R materials. In addition to the
 12 abbreviations previously defined, EAC, M., M&F, MFU, MFUM, MFUML represent a gap

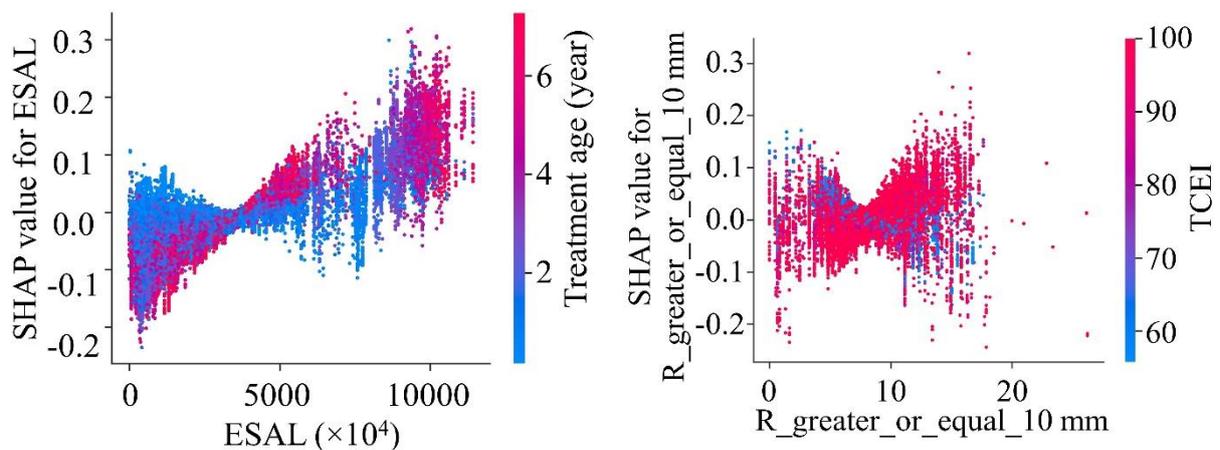
1 graded gradation, SBS modified bitumen, mill and fill, mill and fill the upper asphalt layer,
 2 mill and fill the upper and middle asphalt layer, and mill and fill the entire asphalt layer,
 3 respectively. “Fine M&F” and “thin overlay” means that the milling depth or the overlay
 4 thickness is no more than 2.5 cm. “Fine M&F and overlay” refers to milling the existing
 5 asphalt layer of no more than 2.5 cm, resurfacing it with new asphalt mixtures, and then
 6 adding an asphalt overlay of 2~4 cm on the resurfaced pavement. Its only difference with
 7 “MFU and overlay” is the milling and resurfacing depth. It is apparent from Figure 12 that
 8 seal coating, micro-surfacing, hot-in-place rehabilitation, and fine M&F significantly
 9 decrease the predicted TCEI while the two MFUML treatments lead to the highest SHAP
 10 values. Regarding the other treatments, there is no obvious difference mainly because of the
 11 complicated influence of various factors on the effects of M&R treatments. Therefore, they
 12 need to be discussed case by case.



1 Figure 12. The distributions of SHAP values of M&R treatments.

2 Additionally, an unexpected positive relationship between ESAL and model output
3 was found in Figure 11. To shed further light on this effect, the interaction between ESAL
4 and treatment age was plotted in Figure 13(a). The red points suggest that a larger ESAL
5 would noticeably increase the predicted TCEI when the treatment has served for more than 4
6 years which is somewhat counterintuitive. However, it should be noted that the models in this
7 study are using the TCEI value of the current year to predict that of the following year. Given
8 the dominant importance of the current TCEI, most of the conclusions need to be drawn on a
9 basis of equivalent current TCEI values. Otherwise, the difference in the model output
10 induced by the different values of current TCEI already overwhelms the impact of other
11 features. Meanwhile, ESAL refers to the cumulative traffic loads that have been experienced
12 by the current road segment. Thus, the conclusion should be rephrased as: If the road segment
13 with the same transverse cracking condition has been subjected to more accumulative traffic
14 loads, it is more likely that the predicted TCEI value will be higher. This may be due to the
15 better maintenance effect of the road segment that bear greater traffic loads but reach the
16 same transverse cracking condition. Moreover, this advantage would become more
17 pronounced when the treatment age is over 4 years.

18 The interaction effect between *R_greater_or_equal_10 mm* and current TCEI was
19 also examined. It can be observed from Figure 13(b) that when the current TCEI is lower
20 than 80, a higher ratio of days with 24-hour precipitation greater than or equal to 10 mm will
21 remarkably decline the predicted TCEI value, which might arise from water infiltrating into
22 the cracked asphalt pavement and further exacerbating the cracking.

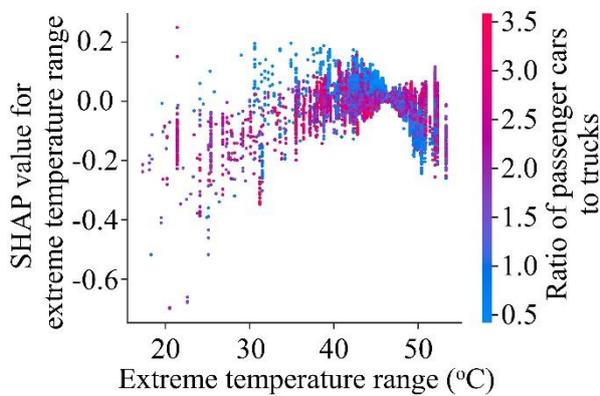


(a) ESAL vs Treatment age

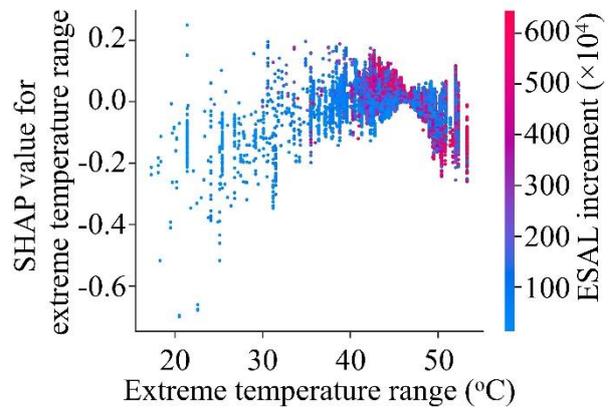
(b) R_greater_or_equal_10 mm/day vs
Current TCEI

1 Figure 13. Visualization of feature interaction effect.

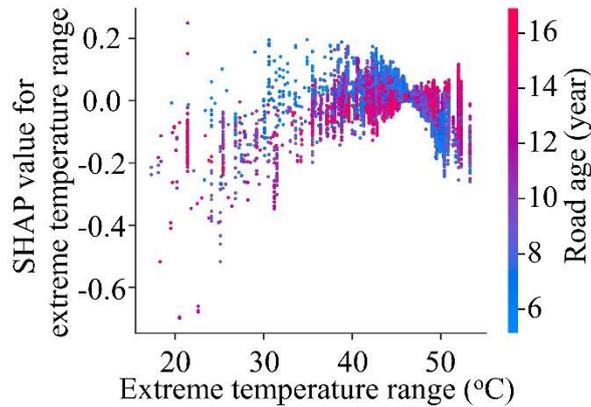
2 Thermal cracking is also a major distress mode in semi-rigid base asphalt pavement
3 caused by the contraction and expansion of the asphalt mixture under extreme temperature
4 changes. Therefore, the top interacting features of extreme temperature range were
5 investigated, as exhibited in Figure 14. Negative correlations could be clearly seen by
6 observing the blue points in Figure 14(a) and (c) and the red points in Figure 14(b). There is a
7 clear trend that with a higher truck ratio, higher ESAL increment or lower road age, the larger
8 extreme temperature range reduces the predicted TCEI. In other words, these conditions
9 make the effect of larger extreme temperature ranges more unfavourable to transverse
10 cracking. It implies that heavy traffic load may aggravate the thermal cracking of asphalt
11 pavement, and thermal cracking could be the main form of early transverse cracks on semi-
12 rigid base asphalt pavement.



(a) Extreme temperature range vs Ratio of
passenger cars to trucks



(b) Extreme temperature range vs ESAL
increment



(c) Extreme temperature range vs Road age

1 Figure 14. Visualization of interacting features of extreme temperature range.

2 5 Conclusions

3 Pavement performance model constitutes an important basis for maintenance decision-
 4 making and budget allocation. While previous studies tended to only focus on model
 5 accuracy, the ability of the model to estimate the uncertainties and interpret the results cannot
 6 be ignored. As a result, this study developed a modelling framework for pavement
 7 performance evolution based on state-of-the-art ML techniques. The importance of feature
 8 selection, uncertainty analysis and model interpretation corresponding to the three modelling
 9 stages: pre-modelling, modelling and post-modelling, was emphasized. BorutaShap algorithm,
 10 Bayesian neural network and SHAP approach were proved to be powerful tools in dealing
 11 with these three tasks.

12 A case study of predicting the pavement transverse cracking was conducted to
 13 demonstrate the application and performance of the proposed framework. It was found that
 14 combining the pavement engineering expertise and the automatic feature selection method
 15 BorutaShap can efficiently determine the optimal features for empirical pavement
 16 performance modelling. A subset of 26 and 27 features that accounts for the pavement age,
 17 material, structure, traffic load, climate condition, maintenance history, etc., was selected for
 18 unmaintained and maintained models, respectively.

19 Two BNN models were established and proved to be accurate in predicting TCEI
 20 values, with the R-square, MAE and RMSE of 0.86, 3.08, 6.21 for unmaintained segments

1 and 0.79, 5.11, 8.50 for maintained segments, respectively. Both model-driven and data-
2 driven uncertainties were estimated. Aleatoric uncertainty was much larger than epistemic
3 uncertainty, indicating that poor data quality played a more dominant role in the total
4 uncertainty. The greater epistemic uncertainty at lower TCEI values also revealed that the
5 developed BNN models have less confidence in predicting lower TCEI values. Thus, it was
6 recommended that practitioners should be careful when using the models to predict possible
7 poor transverse cracking conditions. In addition, further improvement of the predictive power
8 of the model mainly relies on the improvement of the data quality.

9 Interesting findings were also obtained from the interpretation of model output. For
10 example, the explanation of unmaintained model indicated that high mean temperature has
11 greater adverse impact on the transverse cracking of asphalt pavement for road segments with
12 low current TCEI, high road age, or high traffic volume, which may be caused by the
13 accelerated aging of asphalt mixtures in hotter areas. The results of maintained model showed
14 that seal coating, micro-surfacing, hot-in-place rehabilitation, and fine M&F significantly
15 decrease the predicted TCEI while mill & fill the entire asphalt layer leads to the largest
16 model output. It also implied that heavy traffic load may aggravate the thermal cracking of
17 asphalt pavement, and thermal cracking could be the main form of early transverse cracks in
18 semi-rigid base asphalt pavement. Therefore, exploring the additional combined feature effect
19 after accounting for the individual feature effect is necessary and beneficial for practitioners
20 to understand the mechanism of pavement deterioration.

21 Finally, despite the contributions this study has made, there are still opportunities to
22 further extend this research. First, the proposed framework has been only applied to the
23 prediction of transverse cracks. Further studies are needed for its application to other
24 pavement performance indicators. Second, only some features were investigated in this study.
25 Hence, it is better to explore the impact of other features as well so that practitioners can get
26 the whole picture of the impact of different factors on pavement deterioration. Finally, when
27 more data are collected or the quality of data is improved, the model could also be upgraded
28 by allowing it to continue learning.

29 **Acknowledgements**

30 This study was conducted under the support of the Research Institute for Sustainable Urban
31 Development (RISUD) at the Hong Kong Polytechnic University. In addition, the data used

1 in this research were collected from the Pavement Management System in Jiangsu province,
2 China. The engineers and researchers who established the system and collected the data are
3 also acknowledged for their contribution.

4 **Disclosure statement**

5 No potential conflict of interest was reported by the authors.

6 **Reference**

- 7 Amin, S.R. & Amador-Jiménez, L.E., 2017. Backpropagation neural network to estimate
8 pavement performance: Dealing with measurement errors. *Road Materials and*
9 *Pavement Design*, 18 (5), 1218-1238.
- 10 Barber, D. & Bishop, C.M., 1998. Ensemble learning in bayesian neural networks. *Nato ASI*
11 *Series F Computer and Systems Sciences*, 168, 215-238.
- 12 Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D., 2015. Weight uncertainty in
13 neural networks.
- 14 Braswell, E., Saleh, N.F., Elwardany, M., Yousefi Rad, F., Castorena, C., Underwood, B.S. &
15 Kim, Y.R., 2021. Refinement of climate-, depth-, and time-based laboratory aging
16 procedure for asphalt mixtures. *Transportation Research Record*, 2675 (2), 207-218.
- 17 Cai, J., Luo, J., Wang, S. & Yang, S., 2018. Feature selection in machine learning: A new
18 perspective. *Neurocomputing*, 300, 70-79.
- 19 Choi, J.H., Adams, T.M. & Bahia, H.U., 2004. Pavement roughness modeling using back-
20 propagation neural networks. *Computer-Aided Civil and Infrastructure Engineering*,
21 19 (4), 295-303.
- 22 Choi, S. & Do, M., 2020. Development of the road pavement deterioration model based on
23 the deep learning method. *Electronics*, 9 (1), 3.
- 24 Dong, Y., Shao, Y., Li, X., Li, S., Quan, L., Zhang, W. & Du, J., 2019. Forecasting pavement
25 performance with a feature fusion LSTM-BPNN model. *Proceedings of the 28th ACM*
26 *international conference on information and knowledge management*, 1953-1962.

- 1 El-Hakim, A. & El-Badawy, S., 2013. International roughness index prediction for rigid
2 pavements: An artificial neural network application. *Advanced Materials*
3 *ResearchTrans Tech Publ*, 854-860.
- 4 Gal, Y., & Ghahramani, Z., 2015. Dropout as a bayesian approximation: Insights and
5 applications. *In Deep Learning Workshop, ICML* (Vol. 1, p. 2).
- 6 Gal, Y. & Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model
7 uncertainty in deep learning. *33rd International Conference on Machine Learning*.
8 New York, NY, USA
- 9 Gal, Y., Hron, J. & Kendall, A., 2017. Concrete dropout. *arXiv preprint arXiv:1705.07832*.
- 10 Gong, H., Sun, Y., Shu, X. & Huang, B., 2018. Use of random forests regression for
11 predicting IRI of asphalt pavements. *Construction and Building Materials*, 189, 890-
12 897.
- 13 Gong, H., Sun, Y., Hu, W., Polaczyk, P.A. & Huang, B., 2019. Investigating impacts of
14 asphalt mixture properties on pavement performance using LTPP data through
15 random forests. *Construction and Building Materials*, 204, 203-212.
- 16 Guler, S.I. & Madanat, S., 2011. Axle load power for pavement fatigue cracking: Empirical
17 estimation and policy implications. *Transportation research record*, 2225 (1), 21-24.
- 18 Guo, F., Gregory, J. & Kirchain, R., 2020. Incorporating cost uncertainty and path
19 dependence into treatment selection for pavement networks. *Transportation Research*
20 *Part C: Emerging Technologies*, 110, 40-55.
- 21 Han, C., Ma, T., Xu, G., Chen, S., & Huang, R. (2020). Intelligent decision model of road
22 maintenance based on improved weight random forest algorithm. *International*
23 *Journal of Pavement Engineering*, 1-13.
- 24 Hong, F. & Prozzi, J.A., 2006. Estimation of pavement performance deterioration using
25 bayesian approach. *Journal of infrastructure systems*, 12 (2), 77-86.
- 26 Hossain, M., Gopiseti, L. & Miah, M., 2019. International roughness index prediction of
27 flexible pavements using neural networks. *Journal of Transportation Engineering*,
28 *Part B: Pavements*, 145 (1), 04018058.

- 1 Kargah-Ostadi, N. & Stoffels, S.M., 2015. Framework for development and comprehensive
2 comparison of empirical pavement performance models. *Journal of Transportation*
3 *Engineering*, 141 (8), 04015012.
- 4 Karlaftis, A.G. & Badr, A., 2015. Predicting asphalt pavement crack initiation following
5 rehabilitation treatments. *Transportation Research Part C: Emerging Technologies*,
6 55, 510-517.
- 7 Keany, E., 2020. BorutaShap : A wrapper feature selection method which combines the
8 Boruta feature selection algorithm with Shapley values. (Version 1.1). Zenodo.
9 <http://doi.org/10.5281/zenodo.4247618>
- 10 Kendall, A. & Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for
11 computer vision? *arXiv preprint arXiv:1703.04977*.
- 12 Kırbaş, U. & Karaşahin, M., 2016. Performance models for hot mix asphalt pavements in
13 urban roads. *Construction and Building Materials*, 116, 281-288.
- 14 Kobayashi, K., Kaito, K. & Lethanh, N., 2012. A statistical deterioration forecasting method
15 using hidden markov model for infrastructure management. *Transportation Research*
16 *Part B: Methodological*, 46 (4), 544-561.
- 17 Kursa, M.B. & Rudnicki, W.R., 2010. Feature selection with the boruta package. *J Stat Softw*,
18 36 (11), 1-13.
- 19 Li, Q., Xiao, D.X., Wang, K.C., Hall, K.D. & Qiu, Y., 2011. Mechanistic-empirical pavement
20 design guide (mepdg): A bird's-eye view. *Journal of Modern Transportation*, 19 (2),
21 114-133.
- 22 Lundberg, S.M., Erion, G., Chen, H., Degraeve, A., Prutkin, J.M., Nair, B., Katz, R.,
23 Himmelfarb, J., Bansal, N. & Lee, S.-I., 2020. From local explanations to global
24 understanding with explainable ai for trees. *Nature machine intelligence*, 2 (1), 56-67.
- 25 Lundberg, S. & Lee, S.-I., 2017. A unified approach to interpreting model predictions. *arXiv*
26 *preprint arXiv:1705.07874*.
- 27 Madeh Piryonesi, S. & El-Diraby, T.E., 2021. Using machine learning to examine impact of
28 type of performance indicator on flexible pavement deterioration modeling. *Journal of*

- 1 *infrastructure systems*, 27 (2), 4021005.
- 2 Mazari, M. & Rodriguez, D.D., 2016. Prediction of pavement roughness using a hybrid gene
3 expression programming-neural network technique. *Journal of Traffic and*
4 *Transportation Engineering (English Edition)*, 3 (5), 448-455.
- 5 Molnar, C., 2018. A guide for making black box models explainable. URL:
6 <https://christophm.github.io/interpretable-ml-book>.
- 7 Onayev, A. & Sweil, O., 2021. Iri deterioration model for asphalt concrete pavements:
8 Capturing performance improvements over time. *Construction and Building Materials*,
9 271, 121768.
- 10 Osorio-Lird, A., Chamorro, A., Videla, C., Tighe, S. & Torres-Machi, C., 2018. Application
11 of markov chains and monte carlo simulations for developing pavement performance
12 models for urban network management. *Structure and Infrastructure Engineering*, 14
13 (9), 1169-1181.
- 14 Pirayonesi, S.M. & El-Diraby, T.E., 2020. Data analytics in asset management: Cost-effective
15 prediction of the pavement condition index. *Journal of Infrastructure Systems*, 26 (1),
16 04019036.
- 17 Prozzi, J.A. & Madanat, S.M., 2003. Incremental nonlinear model for predicting pavement
18 serviceability. *Journal of transportation Engineering*, 129 (6), 635-641.
- 19 Prozzi, J. & Madanat, S., 2004. Development of pavement performance models by combining
20 experimental and field data. *Journal of Infrastructure Systems*, 10 (1), 9-22.
- 21 Rahman, M.M., Uddin, M.M. & Gassman, S.L., 2017. Pavement performance evaluation
22 models for south carolina. *KSCE Journal of Civil Engineering*, 21 (7), 2695-2706.
- 23 Roberts, R., Inzerillo, L. & Di Mino, G., 2021. Exploiting data analytics and deep learning
24 systems to support pavement maintenance decisions. *Applied Sciences*, 11 (6), 2458.
- 25 Ryu, S., Kwon, Y. & Kim, W.Y., 2019. Uncertainty quantification of molecular property
26 prediction with bayesian neural networks. *arXiv preprint arXiv:1903.08375*.
- 27 Saltelli, A., 1999. Sensitivity analysis: Could better methods be used? *Journal of Geophysical*

- 1 *Research: Atmospheres*, 104 (D3), 3789-3793.
- 2 Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., Li, S. & Wu, Q.,
3 2019. Why so many published sensitivity analyses are false: A systematic review of
4 sensitivity analysis practices. *Environmental modelling & software*, 114, 29-39.
- 5 Shrikumar, A., Greenside, P. & Kundaje, A., Year. Learning important features through
6 propagating activation differences. *International Conference on Machine*
7 *Learning*PMLR, 3145-3153.
- 8 Tabatabaee, N., Ziyadi, M. & Shafahi, Y., 2013. Two-stage support vector classifier and
9 recurrent neural network predictor for pavement performance modeling. *Journal of*
10 *Infrastructure Systems*, 19 (3), 266-274.
- 11 Wittek, P., 2014. Quantum machine learning: What quantum computing means to data
12 mining: *Academic Press*.
- 13 Yao, L., Dong, Q., Jiang, J. & Ni, F., 2019. Establishment of prediction models of asphalt
14 pavement performance based on a novel data calibration method and neural network.
15 *Transportation Research Record*, 2673 (1), 66-82.
- 16 Yao, L., Dong, Q., Jiang, J. & Ni, F., 2020. Deep reinforcement learning for long-term
17 pavement maintenance planning. *Computer - Aided Civil and Infrastructure*
18 *Engineering*, 35 (11), 1230-1245.
- 19 Yehia, A. & Swei, O., 2020. Probabilistic infrastructure performance models: An iterative-
20 methods approach. *Transportation Research Part C: Emerging Technologies*, 111,
21 245-254.
- 22 Zeida, W., Hamad, K., Omar, M., Underwood, B.S., Khalil, M.A. & Karzad, A.S., 2019.
23 Investigation and modelling of asphalt pavement performance in cold regions.
24 *International Journal of Pavement Engineering*, 20 (8), 986-997.
- 25 Zeida, W., Dabous, S.A., Hamad, K., Al-Ruzouq, R. & Khalil, M.A., 2020. Machine
26 learning for pavement performance modelling in warm climate regions. *Arabian*
27 *Journal for Science and Engineering*, 1-19.

1 Zhou, L., Ni, F. & Zhao, Y., 2010. Evaluation method for transverse cracking in asphalt
2 pavements on freeways. *Transportation research record*, 2153 (1), 97-105.

3