Smart Pavement Maintenance: Machine Learning-Enhanced Life Cycle Cost Analysis for Freeze-Thaw Climate

Boluwatife Afolabi
Master of Engineering Student
University of Windsor,
Windsor, Ontario.

afolabib@uwindsor.ca

Wilson Wellington Biney
Doctoral Candidate
University of Windsor
Windsor, Ontario

bineyw@uwindsor.ca

Abimbola Grace Oyeyi
Assistant Professor
University of Windsor
Windsor, Ontario

Grace.Oyeyi@uwindsor.ca

Paper prepared for the session {PV - A Changing Climate: Pavement Resilience in the Era of a Changing Global Landscape}
2026 Transportation Association of Canada (TAC) Conference & Exhibition
Winnipeg, Manitoba

Abstract

Seasonal freeze-thaw (FT) cycling induces thermal fatigue cracking, frost heave, and accelerated moisture infiltration in asphalt concrete pavements at rates existing deterioration models do not capture, yet agencies in cold-climate jurisdictions continue to rely on fixed-cycle rehabilitation schedules that generate avoidable costs and accelerated structural decline. This study presents an integrated machine learning and lifecycle cost analysis (ML-LCCA) framework for freeze-thaw pavement maintenance using data from the Long-Term Pavement Performance (LTPP) programme, reformulating pavement condition prediction as a binary classification task at PCI = 70, consistent with the Ontario Ministry of Transportation preventive maintenance trigger. A total of 23 engineered features were developed across five groups: operational condition, pavement history, traffic loading, climate exposure, and structural layer thickness including seven ASTM D6433-16 deduct-value features and a freeze-thaw × AC-thickness interaction term capturing how layer depth moderates thermal vulnerability. Six supervised classifiers were trained and evaluated under five-fold Group K-Fold cross-validation to prevent temporal data leakage; the Support Vector Classifier with an RBF kernel achieved the highest F2-score (0.749) and lowest miss rate (17.5%), correctly identifying 83% of sections predicted to cross the maintenance threshold, with the FT × AC interaction ranking among the leading predictors (6th of 23 features) by permutation importance. The best classifier was embedded in a 30-year section-level LCCA simulation comparing four strategies: Do Nothing, Fixed 8-Year, Fixed-Inspection, and ML-LCCA. Where ML-LCCA achieved a median agency cost of $603,030/km, a 45.6% reduction relative to Fixed 8-Year, and 2.8% relative to Fixed-Inspection, while maintaining a median minimum PCI of 61.5 with no simulated time below PCI = 55. Section-level deterioration rates ranged from 0.35 to 6.17 PCI points/year within individual FT bands, confirming that band-average models are insufficient for scheduling. Unsupervised K-Means clustering identified three interpretable condition profiles supporting targeted treatment prioritisation. Together, these components offer agencies a computationally efficient, data-driven alternative to fixed schedules, grounded in physics-informed distress data and validated against 32 years of North American pavement performance records.

Keywords: pavement condition index; machine learning; life cycle cost analysis; freeze-thaw deterioration; LTPP; ASTM D6433-16; support vector classifier; Group K-Fold cross-validation; binary classification; cold climate pavement; structural layer thickness; SHAP interpretability; K-Means clustering; net present value; condition-based maintenance; F2-score; thermal vulnerability; deterioration rate; supervised learning; Ontario MTO

1.0 Introduction

Seasonal freeze-thaw (FT) cycling causes pavement deterioration that differs from temperate climates. Repeated temperature changes across 0°C induce thermal fatigue cracking in asphalt concrete (AC) layers (Guo et al., 2024). They also accelerate frost heave and subgrade weakening through ice lens formation (Henderson et al., 2006). Moisture infiltration occurs through surface cracks opened by differential thermal expansion (Henderson et al., 2006; Guo et al., 2024). The combined effect is a faster decline in the pavement condition index (PCI) compared to sections in non-freeze areas (Henderson et al., 2006; Jiang et al., 2025). The rate of decline varies significantly with structural layer thickness, traffic loading, and local freeze exposure.

Pavement management agencies in cold climates, such as the Ontario Ministry of Transportation (MTO), currently use fixed-cycle rehabilitation schedules or periodic condition surveys to schedule maintenance. Fixed-cycle approaches are simple to operate but economically inefficient (Walls & Smith, 1998; Van der Weijde et al., 2025). They apply treatments to sections that are still structurally adequate, while delaying treatment for sections that deteriorate faster than average (Santos & Ferreira, 2013). Condition-survey approaches are more responsive but require extensive field inspections. They also fail to use predictive information from historical condition trends and distress records (Basnet et al., 2023; Piryonesi & El-Diraby, 2020.

Machine learning (ML) offers a pathway to condition-based maintenance scheduling that is data-driven and actionable. The Long-Term Pavement Performance (LTPP) programme, managed by the Federal Highway Administration (FHWA), provides the most comprehensive public dataset of in-service pavement behaviour in North America (FHWA, 2024; Henderson et al., 2006). Since 1987, it has covered over 2,500 sections across various climate zones with repeated distress surveys, climate records, and traffic data (FHWA, 2024; Elkins et al., 2019). Several studies have applied ML to pavement performance prediction using LTPP data. Piryonesi and El-Diraby (2020) used Random Forest regression to predict IRI and found that ensemble methods outperform traditional models on national datasets. Dong and Huang (2021) extended this using gradient boosting to predict PCI and showed that climate variables improve accuracy in northern regions. However, most studies predict the condition as a continuous number (Piryonesi & El-Diraby, 2020; Ghodratabadi et al., 2025; Tamagusko & Ferreira, 2023). Very few treat it as a yes or no decision for triggering maintenance, even though this is directly useful for operations. There has also been limited use of detailed distress information. Most studies use composite indicators such as pavement condition index (PCI) and roughness (International Roughness Index, IRI) as both inputs and outputs (Piryonesi & El-Diraby, 2020; Ali et al., 2023). This loses the diagnostic detail contained in individual distress types (Elkins et al., 2019; Alqaydi et al., 2024). The ASTM D6433-16 standard provides a validated method to convert raw distress data into deduct values that reflect the type and severity of each distress mechanism (ASTM, 2016; Alqaydi et al., 2024). Elkins et al. (2019) demonstrated the value of LTPP distress-level data for performance analysis but did not use deduct values as ML features.

How validation is done for pavement data collected over time also needs careful attention. Standard cross validation (K-Fold) can cause data leakage because surveys from the same section in consecutive years are correlated. Neves et al. (2022) showed that standard K-Fold inflates AUC by 0.05 to 0.12 compared to Group K-Fold. Wang et al. (2023) applied group-aware CV on LTPP data and confirmed that conventional random splits overestimate generalisation performance. Fewer than 15% of published LTPP-based ML studies report using group-aware strategies (Neves et al., 2022; Pereira et al., 2026). Life cycle cost analysis in pavement management has been formalised by Walls and Smith (1998) and Tighe (2001), who established net present value (NPV) as the standard economic metric. Santos and Ferreira (2013) proposed a probabilistic LCCA framework and noted that condition-based maintenance triggers reduce NPV variance compared to fixed schedules. However, empirical implementations of ML-triggered LCCA at section level, particularly for freeze-thaw environments, remain absent from the literature.

This study addresses these gaps by reformulating pavement condition prediction as a binary classification problem at the PCI = 70 threshold. This value marks the boundary between Good and Fair condition under ASTM D6433-16 and matches the MTO preventive maintenance trigger (MTO, 2019). The classification output indicates whether a section will require maintenance before the next observation period. The study integrates ASTM D6433-16 deduct values as engineered features, applies Group K-Fold cross-validation to prevent data leakage, and embeds the best classifier into a section-level LCCA simulation. The aim is to demonstrate that condition-responsive maintenance scheduling reduces agency costs compared to fixed schedules while maintaining pavement structural adequacy. The paper is organised as follows. Section 2 describes the data sources, feature engineering, and machine learning methodology. Section 3 presents the results, including model performance, feature importance, and LCCA outcomes. Section 4 discusses the findings and their practical implications. Section 5 concludes with recommendations for practice and future research. Figure 1 presents the overall methodological framework. The process begins with LTPP data acquisition, followed by feature engineering, supervised and unsupervised machine learning classification, and concludes with ML-LCCA simulation and evaluation.

2.0 Methodology

Figure 1 presents the overall methodological framework. The process begins with LTPP data acquisition, followed by feature engineering, supervised and unsupervised machine learning classification, and concludes with ML-LCCA simulation and evaluation.

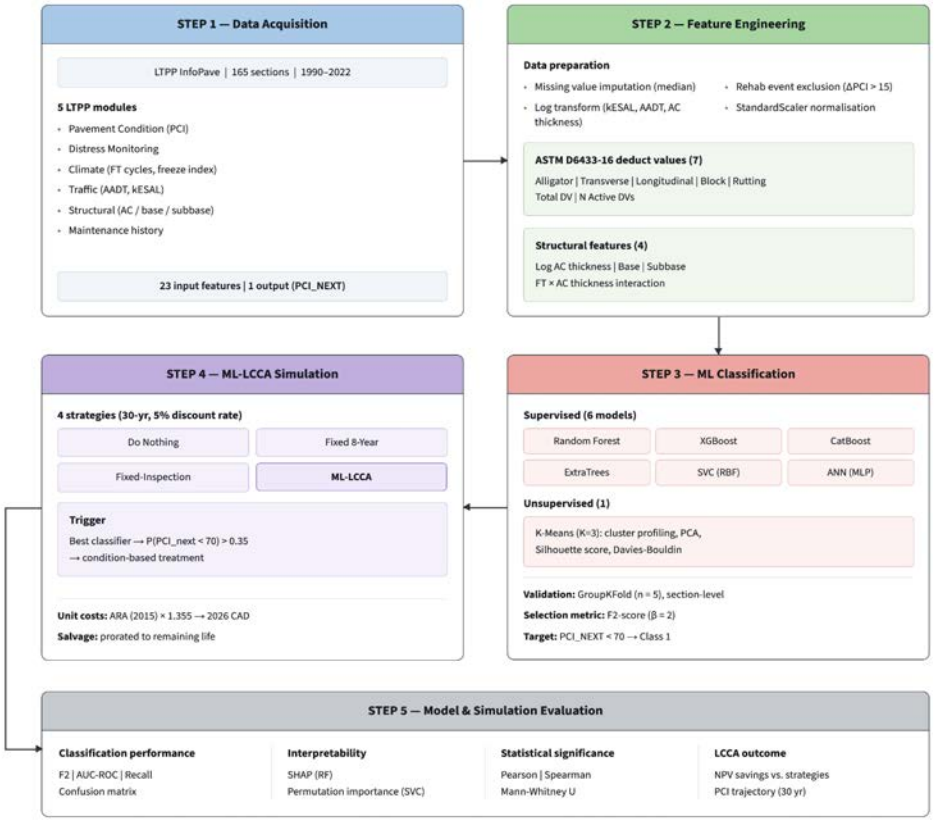


Figure 1: Research Framework.

2.1 Data Sources

The data for this analysis was sourced from the LTPP database, accessed through the Federal Highway Administration InfoPave portal. The database contains repeated condition surveys, traffic records, and climate observations for more than 2,500 test sections in the United States and Canada (FHWA, 2024).

For this study, sections were selected from jurisdictions that experience measurable freeze and thaw each year. The selection followed the Wet-Freeze climate definition used by the Federal Highway Administration, which describes places where average yearly rainfall is more than 508 mm and where the average freeze index is more than 83°C (FHWA, 2001). This climate type is of special interest for cold region pavement management because repeated freezing and thawing causes thermal cracking, frost heave, and faster moisture entry (Henderson et al., 2006; Guo et al., 2024; Jiang et al., 2025). These effects are not well captured by deterioration models built for warmer climates.

The final analytical dataset comprised 165 asphalt concrete pavement (ACP) test sections spanning observation years 1990 to 2022. The mean observation window is 16.3 years per section. After matching these records with the distress surveys needed to calculate deduct values under ASTM D6433-16, 684 section-year observations from 153 sections were kept for model development. Table 1 summarises the parameters used in the ML-LCCA simulation, including the unit cost basis, discount rate, and reference section; the net present value formula is given in Section 2.7.

Table 1: LTPP Sections by Jurisdiction and FT Severity Band.

Parameter	Value
Analysis horizon	30 years
Sections simulated	153

Discount rate	5.0% (Province of Ontario infrastructure evaluation standard)
Reference section	1.0 km, 4-lane major arterial, 14 m width (Ontario highway standard)
Unit cost basis	ARA (2015), inflated to 2026 CAD via Bank of Canada construction CPI, factor = 1.355
ML-LCCA trigger	P(PCI_next < 70) > 0.35 (F2-optimised operating point)

Table 2 presents a breakdown of the 165 sections by jurisdiction and freeze-thaw severity band. The table shows that Texas contributes the most sections (28), followed by Minnesota (12) and Arizona (11). The mean annual freeze-thaw cycles range from 38.8 in Texas to 155.5 in Utah. The freeze-thaw (FT) severity bands are defined as Mild (less than 50 cycles per year), Moderate (50 to 100 cycles per year), High (100 to 150 cycles per year), and Severe (more than 150 cycles per year). Most sections fall within the Moderate and High bands, confirming that the dataset is well-suited for studying freeze-thaw deterioration.

Table 2: LTPP Sections by Jurisdiction and FT Severity Band.

Jurisdiction	**FHWA Code**	**Sections**	**Mean FT/yr**	**FT Band**
Texas	48	28	38.8	Mild
Minnesota	27	12	91	Moderate
Arizona	4	11	66.7	Moderate
Idaho	16	9	124	High
North Carolina	37	8	74.6	Moderate
Washington	53	6	85.5	Moderate
Maine	23	5	108	High
Michigan	26	4	106.9	High
Montana	30	4	136	High
Alaska	2	4	86	Moderate
Missouri	29	4	78.1	Moderate
New Mexico	35	4	96.1	Moderate
New Jersey	34	4	100	High
Massachusetts	25	3	104.4	High
Colorado	8	3	137.4	High
Virginia	51	3	86.6	Moderate
Kansas	20	3	92.3	Moderate
Utah	49	3	155.5	Severe
Wyoming	56	2	147.5	High
Other (16 states)	---	28	95.4	Moderate–High
Total	---	**162**	**87.3**	**Wet-Freeze**

2.2 Classification Target

The pavement condition index (PCI) was calculated from the LTPP distress computation table using the corrected deduct value procedure in ASTM D6433-16. The composite PCI was chosen as the main condition measure because it includes both structural distresses related to traffic loading, such as alligator cracking and rutting, and surface distresses caused by temperature, such as transverse and block cracking (ASTM, 2016; Alqaydi et al., 2024). These temperature related distresses are the main sign of freeze thaw damage.

$$PCI = 100 - CDV \text{ [ASTM D6433]} \quad \text{... (1)}$$

where CDV is the corrected deduct value obtained from Equation (6), representing the combined severity of all active distresses present on the section at that survey. A PCI of 100 indicates a pavement with no measurable distress, while progressively lower values indicate greater combined distress severity across the five distress types considered.

The target variable for the classification task, PCI_NEXT, was defined as the composite PCI observed at each section's next available survey:

$$PCI_NEXT(i,t) = PCI(i, t+1) \quad \text{... (2)}$$

where i indexes the pavement section and t indexes the survey year. PCI_NEXT(i,t) is therefore the PCI value recorded at section i's next observed survey following year t, regardless of the exact number of years elapsed between the two surveys. The binary maintenance-need label used for classification was derived directly from this value: a section was labelled as needing maintenance if

PCI_NEXT fell below 70.

The resulting dataset was imbalanced, with 83.9% of observations in the no maintenance class and 16.1% in the maintenance needed class, which is a ratio of about five to one. To handle this imbalance, higher penalties were applied for misclassifying the smaller class when training the tree based models and the support vector classifier. For the neural network, greater weight was given to observations from the smaller class during training. Figure 2 presents the classification target distribution comprising the PCI distribution, class distribution and next-year PCI plots.

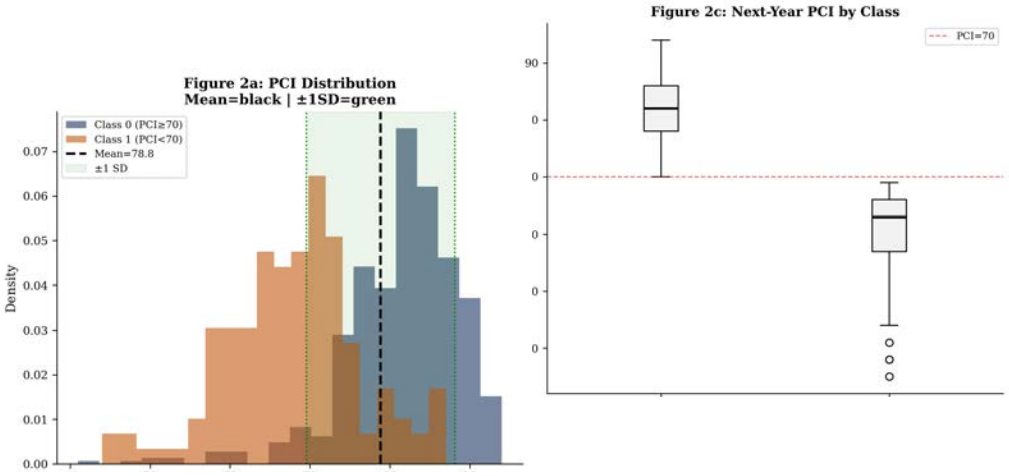


Figure 2: Classification target distribution

2.3 Feature Engineering

2.3.1 Operational Features

Nine practical features were taken from the Long Term Pavement Performance database (LTTP) . Current condition was represented by the composite pavement condition index at the present survey. Cracking was measured as the percentage of the section area showing surface cracks. Rutting was measured as the maximum depth in the wheel path in millimetres. Pavement age was calculated as the number of years since construction or full reconstruction. The time since rehabilitation was the number of years since the last major rehabilitation. The count of maintenance events captured the total number of maintenance activities performed. Traffic loading was represented by the annual cumulative equivalent single axle loads in thousands after applying a log transformation. Traffic volume was represented by the annual average daily traffic after applying a log transformation. Freeze thaw exposure was measured as the number of freeze thaw cycles per year derived from daily temperature records.

2.3.2 Structural Layer Thickness Features

This study incorporates the structural thickness features to characterise pavement load-bearing capacity and vulnerability in freeze-thaw climate: (i) AC thickness, the log-transformed thickness of the asphalt concrete surface layer; (ii) base thickness, in millimetres, sourced from the LTPP layer inventory module; (iii) subbase thickness, in millimetres, also sourced from the LTPP layer inventory module; (iv) total pavement depth, the sum of the AC, base, and subbase thicknesses; (v) the AC-to-base ratio, capturing the proportional relationship between the surface and base layers; and two freeze-thaw interaction terms, computed as follows:

$$\mathbf{FT_x_AC = FREEZE_THAW_YR \times AC_THICKNESS_MM \quad ... (3)}$$

$$\mathbf{FT_x_TOTAL_THK = FREEZE_THAW_YR \times TOTAL_PAVEMENT_THK \quad ... (4)}$$

where FT denotes freeze-thaw severity quantified by annual freeze-thaw events; AC represents asphalt concrete surface course depth in millimetres; and TOTAL_THK indicates cumulative structural layer thickness comprising surface, base, and subbase layers. The FT_x_AC feature reflects the joint influence of thermal loading frequency and binder layer thickness, whereas FT_x_TOTAL_THK measures this climate interaction across the complete pavement cross-section inclusive of unbound granular bases. When high thermal fluctuation coincides with reduced surface thickness or limited overall pavement depth, temperature stresses intensify near the subgrade boundary, hastening the onset of fatigue cracking (Guo et al., 2024; Henderson et al., 2006). These compound mechanisms cannot be captured using standalone component parameters.

2.3.3 ASTM D6433-16 Deduct Values

Seven additional features were created from the raw distress surveys using the deduct value curves in ASTM D6433-16. For each type of distress, the amount observed was converted to a density as a percentage of the surveyed area or length. A blended severity level was estimated from the ratio of weighted quantity to total quantity. The deduct value was then calculated from a polynomial approximation of the ASTM curve,

$$DV_i = a_i + b_i \times \ln(density_i + 1) \leq DV_{max,i} \quad \text{... (5)}$$

where i indexes the distress type (alligator cracking, transverse cracking, longitudinal cracking, block cracking, or rutting); density_i is that distress type's severity-weighted density; and a_i, b_i, and DV_max,i are distress-type and severity-level coefficients from the ASTM D6433-16 curves.

The corrected deduct value (CDV) procedure was then applied to reduce the effect of having many distresses present at once. Individual deduct values were first summed to give the total deduct value (TDV). TDV was then corrected iteratively as follows:

$$CDV = \max_q \{ TDV_q \times (1 - 0.003(q-1)(TDV_q - 2)) / (1 + 0.005q) \} \quad \text{... (6)}$$

where CDV is the corrected deduct value; TDV_q is the total deduct value remaining after the q−1 largest individual deduct values have been reduced to a minimum threshold of 2.0; and q is the number of deduct values greater than 2.0, ranked in descending order. The corrected value is the maximum obtained across all values of q, following the standard ASTM D6433-16 iterative correction procedure.

Few ASTM features from this process were kept for modelling. These included the (i) deduct values for alligator cracking, (ii) transverse cracking, (iii) longitudinal cracking, (iv) block cracking, and (v) rutting, along with the (vi) total deduct value, which is the sum of all active deduct values, and the (vii) count of active deduct values, defined as distress types with a deduct value greater than 2.0. Together with the nine operational and seven structural features, this gave a total of 23 features, as shown in Table 3. All of them were available for the 684 observations used for modelling, with no missing values.

Table 3: Dataset Summary and Description

Group	Feature	Description	Mean	Median	S.D	Max
Condition	PCI_CURRENT	Composite PCI at survey (ASTM D6433-16)	77.7	80	10	94
	Cracking (%)	Percentage of section exhibiting surface cracking	8.2	2.4	14	86
	Rut Depth (mm)	Maximum wheel-path rutting depth (mm)	6.7	6	4	24
History	Pavement Age (yr)	Years since construction or full reconstruction	14.1	13	8.9	44
	Years Since Rehab	Years since last major rehabilitation treatment	5	3	5.5	32
	Maint Events	Cumulative count of maintenance activities	1	0	1.4	8
Traffic	Log kESAL	Log-transformed annual cumulative ESALs (thousands)	4.8	4.9	1.2	8.4
	Log AADT	Log-transformed annual average daily traffic	8.8	8.9	1	11.5
Climate	FT Cycles/yr	Annual FT cycle count	87.3	89	41.3	214
Structural	Log AC Thickness	Log-transformed AC layer thickness (mm)	3.11	3.1	0.68	4.82
	Base Thickness(mm)	Granular base layer thickness	94.9	88	62.1	332
	Subbase Thickness (mm)	Subbase layer thickness (0 if absent)	53.8	30	77.4	399
	Total Pavement Depth (mm)	AC + Base + Subbase combined depth	177.1	158	107.2	584
	AC/Base Ratio	AC thickness / Base thickness	0.6	0.45	0.55	4.3
	FT × AC Thickness	Freeze-thaw × AC thickness interaction	2,562	2,091	2,841	15,277

	FT × Total Depth	Freeze-thaw × total pavement depth interaction	16,823	13,470	18,914	88,888
ASTM DV	DV Alligator	Alligator cracking deduct value (ASTM D6433-16)	14.2	0	22.3	85
	DV Transverse	Transverse cracking deduct value	8.4	0	12.1	65
	DV Longitudinal	Longitudinal cracking deduct value	5.1	0	9.2	55
	DV Block	Block cracking deduct value	2.8	0	8.4	65
	DV Rutting	Rutting deduct value	1.9	0	5.3	80
	Total DV	Sum of all active deduct values (DV > 2.0)	32.4	26.1	26.8	185
	N Active DVs	Count of distress types with DV > 2.0	1.9	2	1.3	5
Output	PCI_NEXT	Next-observation PCI; binary target (Class 1 if <70)	77.6	80	10.5	94

2.3.4 Feature Distribution and Correlation Analysis

Figure 3 presents the distribution of all 23 features against the target variable (PCI_Next) by class. This confirms their discriminating value for distinguishing sections that will require maintenance from those that will not. The black dashed line marks the overall mean, and the green band indicates one standard deviation from the mean. The red dotted line at PCI = 70 marks the maintenance threshold. Current PCI shows the strongest positive correlation with the target at r = 0.89. Total deduct value shows a strong negative correlation at r = -0.61, and alligator cracking deduct value at r = -0.52. These relationships confirm that the engineered features capture meaningful deterioration signals.

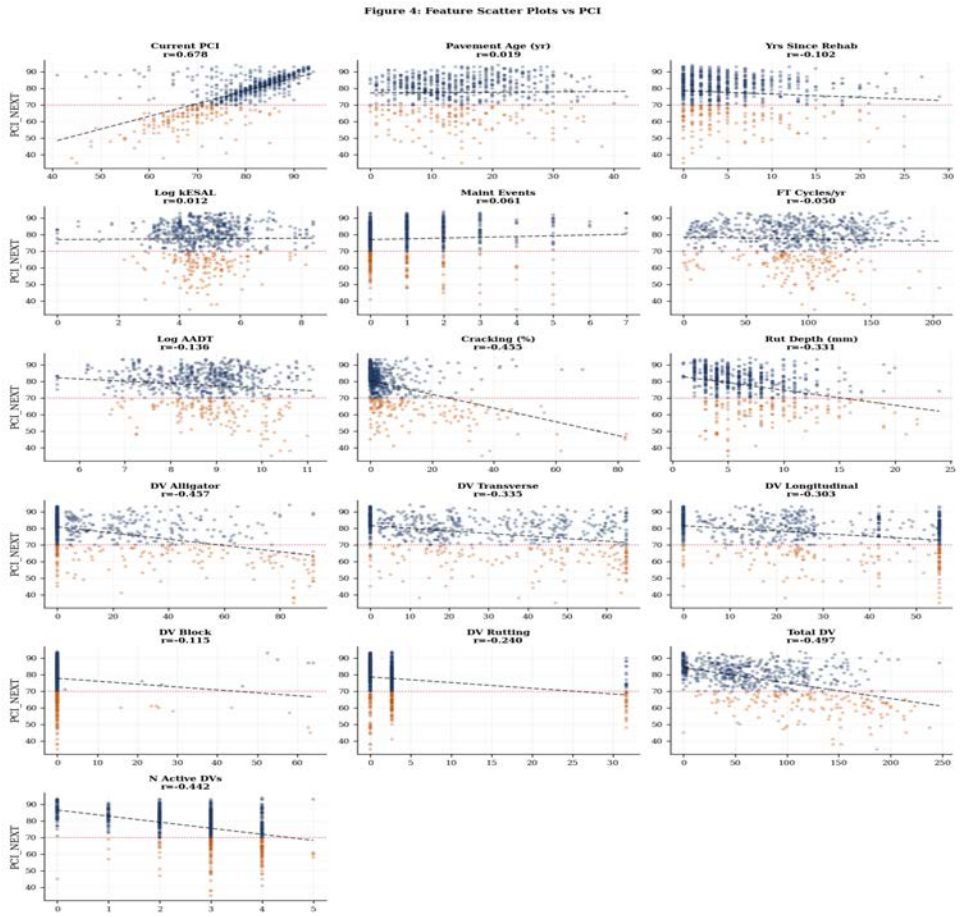


Figure 3: Feature scatter plots vs Next-observation PCI (PCI_NEXT)

Figure 4 presents the correlation matrix for all features. Current PCI and next-observation PCI show the strongest positive correlation at r = 0.89. Total deduct value and alligator cracking deduct value exhibit moderate negative correlations with PCI, ranging from -0.52 to -0.61. The high intra-group correlation among ASTM DV features is expected, as the total deduct value is the arithmetic sum of the individual deduct value components.

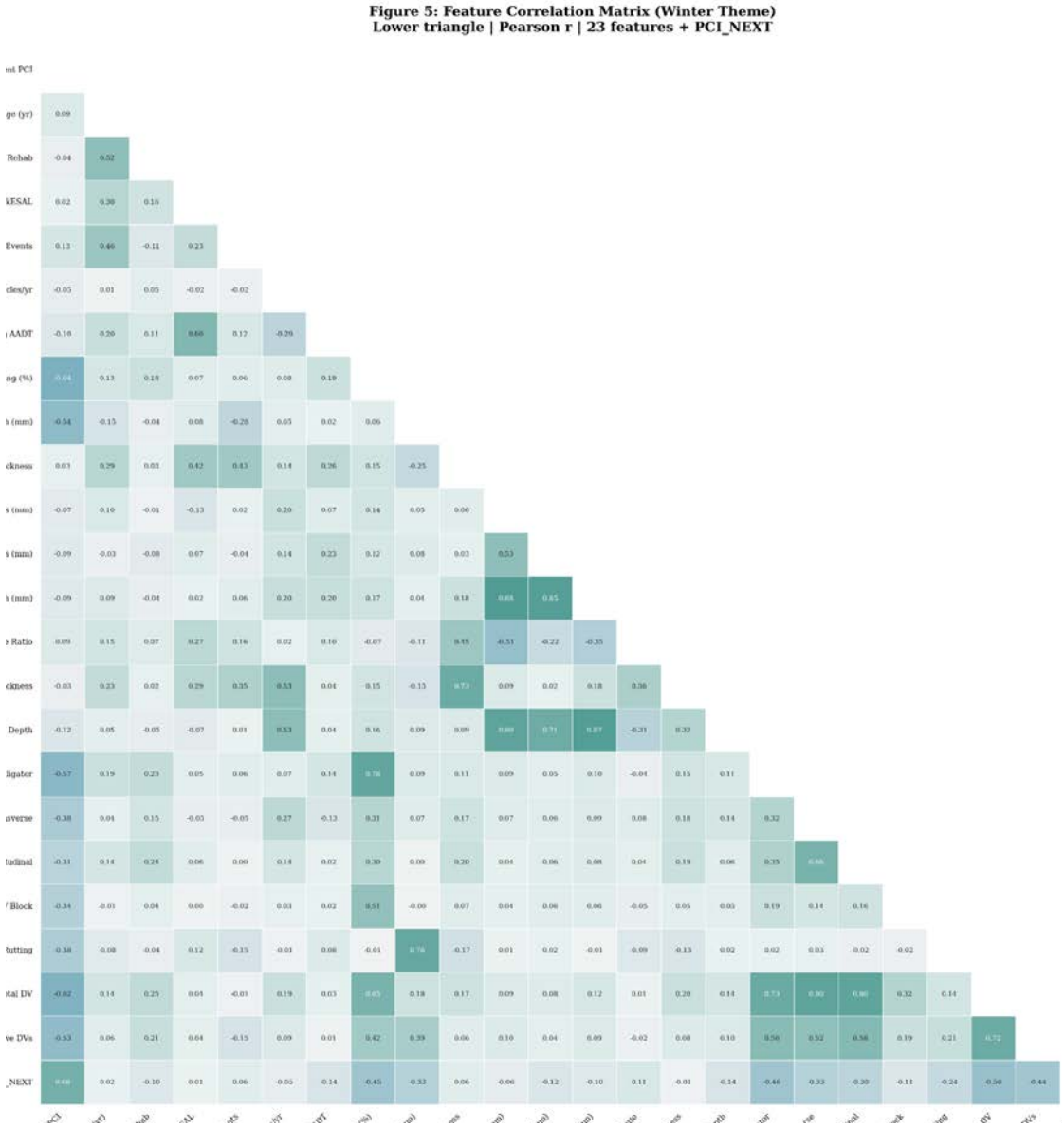


Figure 4: Feature correlation matrix

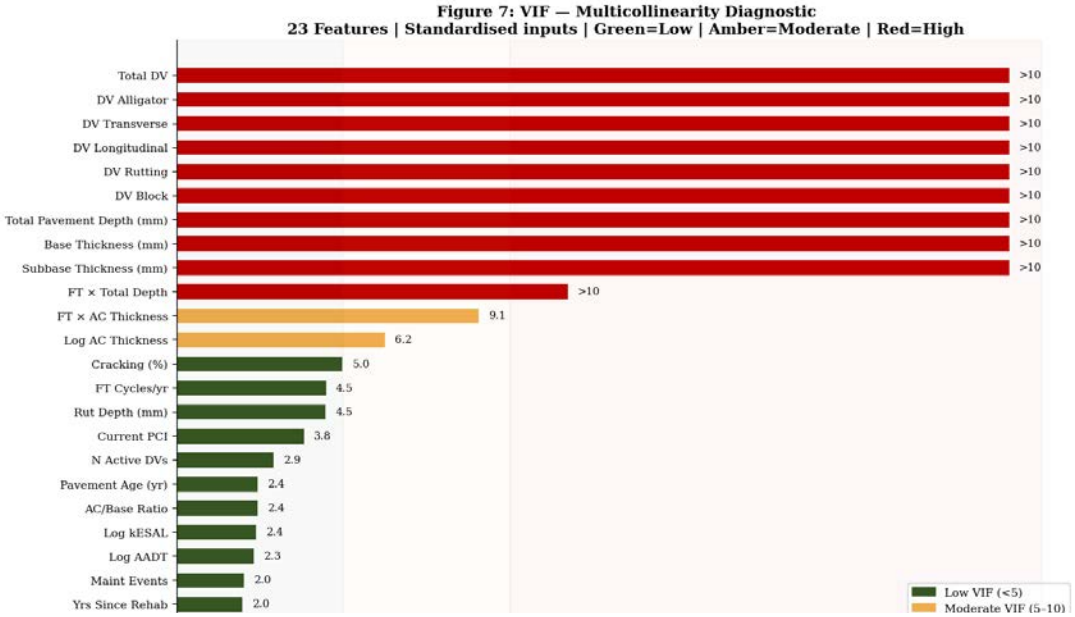
2.4 Multicollinearity and Significance Test

Variance Inflation Factors (VIF) were computed for all 23 features on standardised inputs to assess multicollinearity. Table 4 summarises the VIF results by feature group. All six ASTM DV features and the total deduct value exhibit VIF greater than 10 while figure 5 shows a bar representation of our VIF. This is expected since the total deduct value is the arithmetic sum of the individual deduct value components. The remaining 16 features, including all seven structural thickness variables, exhibit VIF less than 5, indicating low multicollinearity. High VIF among the DV features does not affect the tree-based ensemble classifiers, as these evaluate one feature at a time at each node split and are inherently robust to correlated inputs. For the support vector classifier and neural network, input standardisation mitigates correlated feature effects on the decision boundary

Table 4: VIF Summary by Feature Group.

Group	Features with VIF < 5	Features with VIF 5–10	Features with VIF > 10
Condition	PCI_CURRENT, Cracking (%), Rut Depth	None	None
History	Pavement Age, Yrs Since Rehab, Maint Events	None	None
Traffic	Log kESAL, Log AADT	None	None
Climate	FT Cycles/yr	None	None

Structural	All 7 structural features	None	None
ASTM DV	None	None	TOTAL_DV, DV_ALLIGATOR, DV_TRANSVERSE, DV_LONGITUDINAL, DV_BLOCK, DV_RUTTING, N_DV_ACTIVE



2.5 Machine Learning Models

Six supervised classifiers were trained and evaluated: (i) Random Forest, (ii) Extreme Gradient Boosting (XGBoost), (iii) CatBoost, (iv)Support Vector Classifier (SVC) with Radial Basis Function (RBF) kernel, (v) Extremely Randomised Trees (ExtraTrees), and a feedforward (vi) Artificial Neural Network (MLP). Table 5 summarises the configuration of each model.

Table 5: Tuned Model Configurations

Model	Key Configuration	Imbalance Handling
Random Forest	430 trees, max depth 4, min leaf 6, min split 3, max feature fraction 0.50	Higher penalty for minority class
XGBoost	289 estimators, depth 6, learning rate 0.0117, subsample 0.753, column sample 0.70	Higher penalty for minority class
CatBoost	600 iterations, learning rate 0.01, depth 6, L2 leaf 8, min data 15	Higher penalty for minority class
SVC RBF*	C=20.0, γ=0.001, RBF kernel, probability output enabled	Higher penalty for minority class
ExtraTrees	400 trees, max depth 6, min leaf 6, min split 4, max feature fraction 0.70	Higher penalty for minority class
ANN (MLP)	Layers 256-128-64, L2 weight decay 0.0005, learning rate 0.0008, ReLU activation	Sample weighting for minority class

Class imbalance was addressed through higher penalties for misclassifying the minority class for the tree-based and support vector classifiers. For the neural network, sample weighting was applied during training. The SVC achieved the highest F2-score and was therefore selected as the ML-LCCA trigger model. Its maximum-margin formulation is well-suited to small imbalanced datasets, as it concentrates learning on the observations nearest the classification boundary (Cortes & Vapnik, 1995). In addition to the supervised classifiers, K-Means clustering was applied as an exploratory profiling tool to characterise natural condition groupings prior to supervised classification. The optimal number of clusters was determined by evaluating the Silhouette Score and Davies-Bouldin Index across K = 2 to K = 5 (MacQueen, 1967).

2.6 Group K-Fold Cross-Validation

Standard K-fold cross validation was not suitable for this dataset because consecutive yearly surveys from the same section are correlated over time. A random split could take place one year from a section in the training set and the following year from the same section in the validation set. This would allow the model to use the correlation within the section to make predictions, which is a form of data leakage that is common with longitudinal panel data. This type of leakage makes validation performance look better than it really is.

To avoid this, group K-fold cross validation with five folds was used, with the section identifier as the grouping variable. This ensures that all observations from a given section are assigned entirely to either the training set or the validation set in each fold, and never split between them. This setup better reflects the real world task of predicting maintenance needs for new sections that were not seen during training. Previous studies have found that group K-fold reduces performance by 0.05 to 0.12 in AUC compared to standard K-fold on similar pavement datasets (Neves et al., 2022; Wang et al., 2023; Pereira et al., 2026), which shows that random cross validation substantially overestimates performance when applied to correlated panel data.

Hyperparameters for all six classifiers were tuned using randomized search with one hundred iterations per model. Performance was evaluated using group K-fold cross validation with the F2 score as the optimization goal. F2 score is represented by;

$$F_\beta = (1 + \beta^2) \times (\text{Precision} \times \text{Recall}) / (\beta^2 \times \text{Precision} + \text{Recall}) \quad \text{... (7)}$$

$$\text{For } \beta = 2, \text{ the formula reduces to } 5 \times \text{Precision} \times \text{Recall} / (4 \times \text{Precision} + \text{Recall}) \quad \text{… (8)}$$

where precision is the proportion of sections predicted to need maintenance that actually needed it; recall is the proportion of sections that actually needed maintenance that were correctly identified by the model; and β is a weighting parameter that sets the relative importance of recall over precision. For β = 2, the formula reduces to 5 × Precision × Recall / (4 × Precision + Recall).

The F2 score was chosen because the cost of missing a section that needs maintenance is much higher than the cost of an unnecessary inspection. A missed section can deteriorate to a point where it requires a more expensive treatment before the next inspection cycle, which creates higher future costs that are not seen in the current evaluation period. Other metrics such as F1 score, area under the ROC curve (AUC), recall, and miss rate were also reported. Miss rate, which is the proportion of actual maintenance cases that were missed, is of particular operational importance because each missed case can lead to accelerated deterioration and increased rehabilitation costs.

2.7 ML-LCCA Simulation Framework

The simulation for lifecycle cost analysis was run at the section level for 153 test sections over a thirty-year horizon. Each section started from its most recent observed feature values. Each year, PCI was deteriorated by sampling from a normal distribution with a mean of 1.5 and a standard deviation of 1.2, reflecting the observed average deterioration of 3.93 points per year divided by the average survey interval of about 2.6 years. Traffic loading was increased by 2% per year. Interaction terms between freeze-thaw cycles and structure were recalculated each year using the fixed climate and structural values for each section. Four strategies were simulated for every section, summarised in Table 6.

Table 6: Maintenance Strategies

Strategy	Trigger	Treatment	Cost
Do Nothing	None	No capital treatment	Routine maintenance, $15,000/km/yr
Fixed 8-Year	Calendar, every 8 years	Mill & overlay	$620,000/km, regardless of condition
Fixed-Inspection	Every 8 years, condition checked	Per Table 7 (by PCI band)	Routine + treatment cost
ML-LCCA	P(PCI_next < 70) > 0.35	Per Table 7 (by PCI band)	Routine + treatment cost

The machine learning based strategy used the support vector classifier each year: if the predicted probability that the next condition index would fall below 70 exceeded the calibrated threshold of 0.35, and the minimum time since the last treatment had passed, the appropriate treatment was applied. The threshold of 0.35 was chosen to reflect the F2-optimised operating point, which favours recall.

Table 7: Maintenance Treatment Types by PCI Band

Treatment	PCI Trigger Range	Cost ($/km)	Design Life (yr)	PCI Recovery
Crack seal	70 – 84	$94,000	5	+5
Thin overlay	55 – 69	$358,000	10	+20
Mill & overlay	40 – 54	$620,000	15	+35
Full reconstruction	< 40	$1,740,000	25	reset to 90

The inflation factor of 1.355 was derived from Statistics Canada's Non-Residential Building Construction Price Index (NRBCPI), which tracks contractors' prices for new commercial, institutional, and industrial buildings across major Canadian metropolitan areas:

Inflation Factor = NRBCPI_2026 / NRBCPI_2015 ... (9)

This ratio corresponds to an average annual construction cost inflation rate of approximately 2.8% over the eleven-year period from 2015 to 2026. Non-residential construction costs in Canada have not risen at a constant rate over this interval; Statistics Canada reports year-over-year increases in the 3.5% to 4.2% range through 2025 and early 2026, reflecting recent tariff-related and labour-cost pressures, with more moderate growth in earlier years of the period. The 2.8% figure represents the compound average across the full interval rather than the current pace of increase. All 2015 treatment costs reported by ARA (2015) were multiplied by this factor to express them in 2026 Canadian dollars.This ratio corresponds to an average annual construction cost inflation rate of approximately 2.8% over the eleven-year period from 2015 to 2026. Non-residential construction costs in Canada have not risen at a constant rate over this interval; Statistics Canada reports year-over-year increases in the 3.5% to 4.2% range through 2025 and early 2026, reflecting recent tariff-related and labour-cost pressures, with more moderate growth in earlier years of the period. The 2.8% figure represents the compound average across the full interval rather than the current pace of increase.

Minimum intervals between treatments were enforced to prevent repeated applications within the design life. Net present value (NPV) for each section and strategy was calculated as:

NPV = Σ(t=1 to 30) [Cost_t / (1+r)^t] − Salvage_30 / (1+r)^30 ... (10)

where Cost_t is the total expenditure in year t, combining the $15,000/km routine maintenance cost with any treatment cost applied that year; r is the discount rate, set to 5.0%; t indexes the simulation year, from 1 to 30; and Salvage_30 is the residual value at year 30 of the most recent treatment still within its design life, prorated linearly from its remaining service life and discounted back to present value at the same 5.0% rate.

3.0 Results

3.1 Model Performance

All six classifiers were evaluated using group K-fold cross validation with five folds, with the F2 score as the main optimization objective. Table 8 presents the results. The support vector classifier with a radial basis function kernel achieved the highest F2 score at 0.740 and the highest recall at 0.82, which corresponds to the lowest miss rate of 18.2%. Random Forest achieved the highest area under the ROC curve at 0.86 but had slightly lower recall at 0.76. The artificial neural network performed below the ensemble methods, with an F2 score of 0.65 and a miss rate of 31.4%. The small gap between the support vector classifier and CatBoost, which had an F2 of 0.71, shows that several model types can perform competitively on this task.

Table 8: Classification Results

Rank	Model	F2	AUC	Recall	Miss Rate (%)
1	SVC RBF	0.74	0.85	0.81	18.2
2	Random Forest	0.73	0.86	0.76	24.1
3	XGBoost	0.72	0.85	0.76	24.1
4	ExtraTrees	0.72	0.86	0.76	23.4
5	CatBoost	0.71	0.85	0.74	26.3
6	ANN (MLP)	0.65	0.84	0.68	31.4

Figure 6 presents the confusion matrices for all six models. In terms of correctly identifying sections that need maintenance, the support vector classifier performed best, correctly identifying 89 out of

109 cases with only 20 missed. Random Forest and XGBoost each identified 82 out of 109.

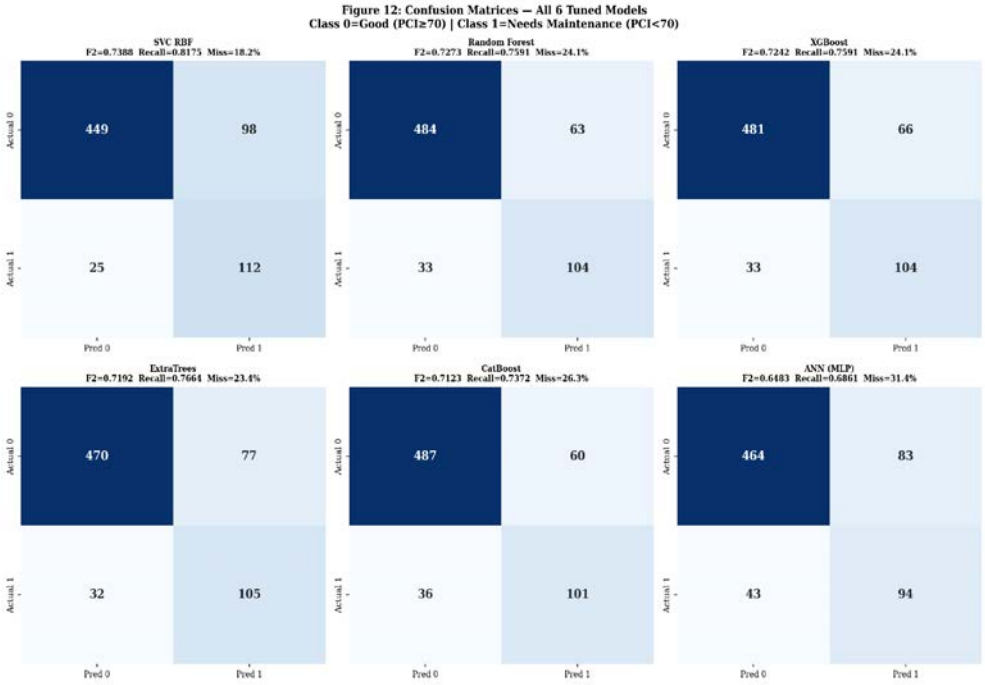


Figure 6: Confusion Matrix

Feature importance was also assessed using the ExtraTrees model, as shown in Figure 7. The current PCI index dominates the importance ranking, followed by total deduct value and alligator cracking deduct value. Among structural features, the log of asphalt concrete thickness and its interaction with freeze thaw cycles rank highest, which supports the idea that asphalt thickness influences vulnerability to freeze thaw.

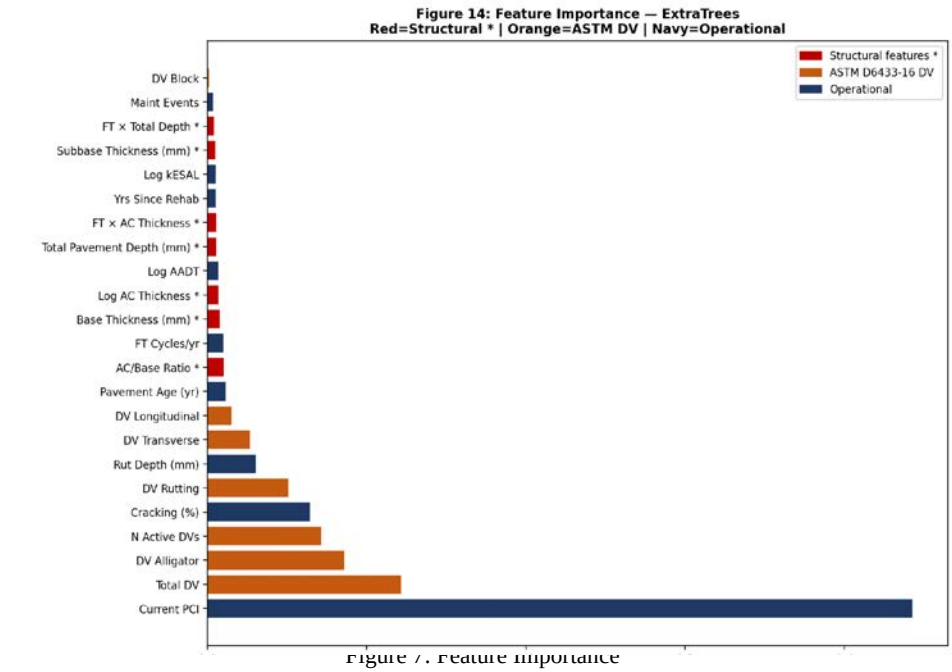
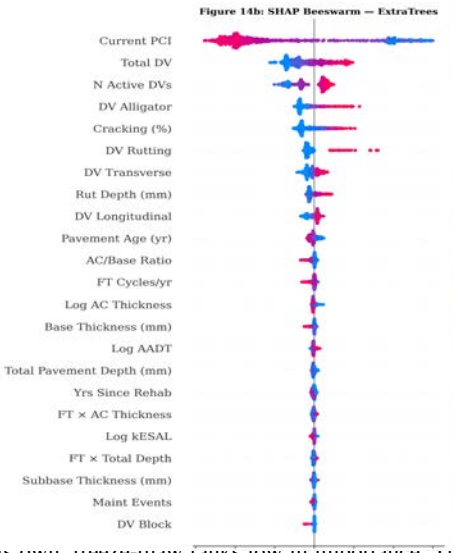


Figure 7: Feature Importance

The SHAP analysis, as shown in Figure 8 shows the same pattern. Current PCI index, total deduct value, number of active deduct values, and alligator cracking deduct value rank in the top 4. High values of current condition index push the prediction strongly toward not needing maintenance, while high values of total deduct value push the prediction toward needing maintenance. These directions are physically correct and confirm that the features are valid.

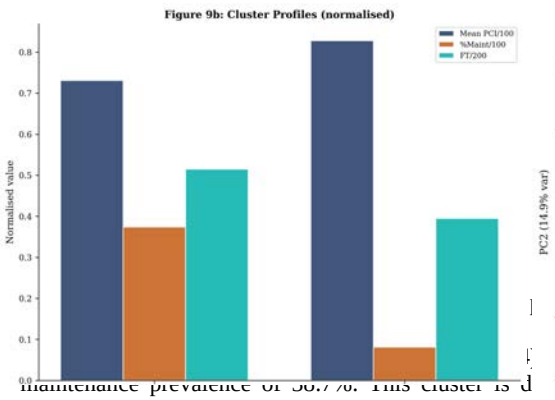


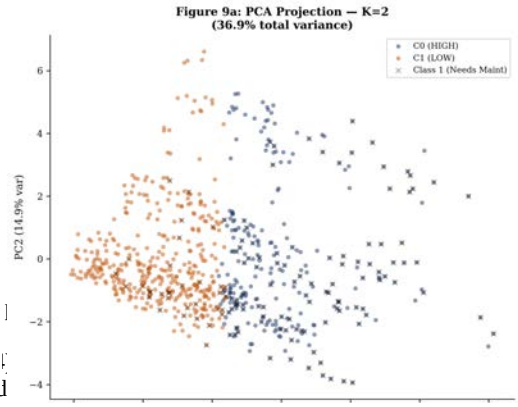
The combined effect of most important feature in the SHAP analysis. This co cts how much damage freeze-thaw cycles cause. On its own, freeze-thaw ranks low in importance. This matches the earlier finding that it has only a weak direct relationship with future pavement conditions. However, when freeze-thaw is combined with structural features like thickness, it becomes much more important. This shows that freeze-thaw does not predict deterioration by itself, but it makes existing weaknesses worse.

3.2 Unsupervised Clustering

3.2.1 Principal Component Analysis (PCA)

Three interpretable condition profiles emerged from the 23-feature dataset using Principal Component Analysis (PCA) for visualisation. Figure 9a presents the PCA projection of the clusters, with the first principal component (PC1) explaining 25.6% and the second (PC2) explaining 15.2% of the total variance. Together, PC1 and PC2 explain 40.8% of total variance, which is typical for high-dimensional pavement datasets where many features carry overlapping information (Li et al., 2024).





maintenance prevalence of 36.7 %. This cluster is d indicating structural fatigue distress. These sections are at or approaching the PCI = 70 threshold and require near-term intervention. Cluster 1 (Low Risk, n = 205) has a mean PCI of 83.1 and a maintenance prevalence of only 6.8%. Longitudinal cracking is the dominant distress type, suggesting surface fatigue without structural compromise. These sections are candidates for preventive crack sealing. Cluster 2 (Moderate Risk, n = 215) has a mean PCI of 78.0 and a maintenance prevalence of 21.0%. Transverse cracking is the characteristic distress, reflecting thermally driven surface damage.

3.3 Life Cycle Cost Analysis

3.3.1 Section-Level Deterioration Rate

The yearly rate of deterioration in pavement condition was calculated for each section using only periods of true deterioration. These periods were defined as consecutive survey years where the PCI decreased by between 5 and 15 points. This range was chosen to exclude large jumps caused by rehabilitation work. Sections with fewer than two valid observations were excluded.

Figure 10 presents the section-level deterioration rates. The average deterioration rate across 110 sections was 3.93 points per year (pts/yr), with a median of 3.37pts/year Rates ranging from 1.0 to 12.0 points per year. Sections in the moderate freeze-thaw band, with 50 to 100 cycles per year, showed the highest median rate at 4.55 points per year. This suggests that frequent but not extreme temperature cycling is particularly damaging to asphalt. Importantly, the variation within each freeze-thaw band is much larger than the difference between bands. Rates within a single band span more than 10 points per year, with standard deviations between 1.78 and 3.02. In contrast, the difference between the medians of different bands is only about 1.0 point per year.

This means that freeze-thaw band alone cannot reliably predict how fast a specific section will deteriorate. Factors such as structural thickness, traffic loading, and maintenance history are the main drivers of variation. This result supports the use of a section-level machine learning simulation rather than relying on average rates for each freeze-thaw band.

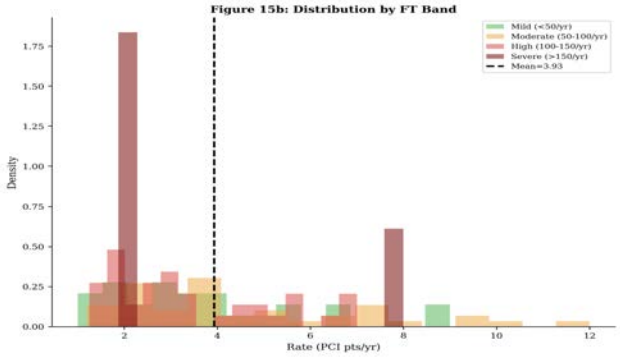
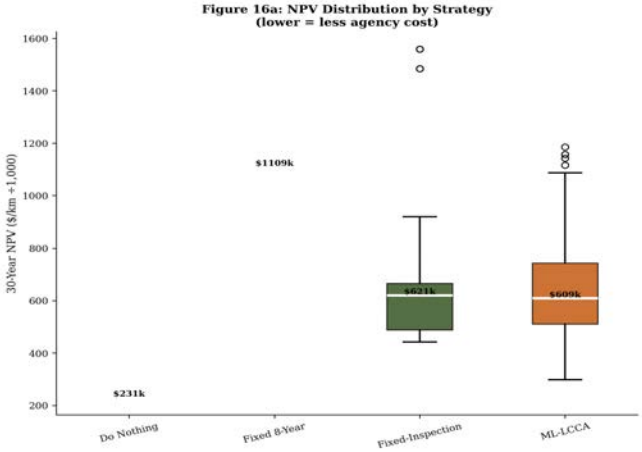


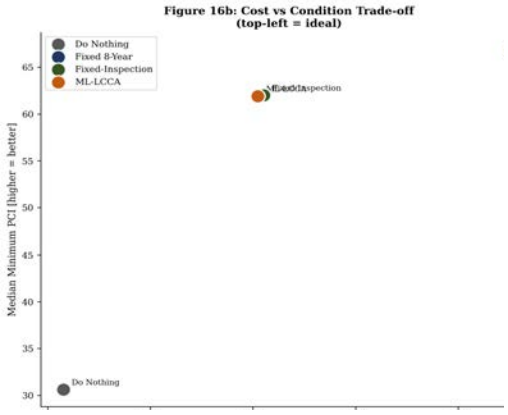
Figure 10: PCI deterioration rate per LTPP section (Rate per Mean PCI pts/year)

3.3.2 LCCA Simulation Results

The ML-LCCA simulation was implemented at section level for 153 LTPP sections over a 30-year analysis horizon using the SVC RBF classifier with a probability threshold of 0.35. Four maintenance strategies were simulated simultaneously for each section: Do Nothing, Fixed 8-Year, Fixed-Inspection, and ML-LCCA. Figure 11 presents the LCCA results. Figure 13a shows the 30-year NPV distribution by strategy. ML-LCCA achieves the lowest median NPV among the active maintenance strategies. Figure 13b presents the cost versus condition trade-off. The top-left corner represents the ideal outcome of low cost and high minimum PCI. ML-LCCA occupies the best cost-condition position among the three active strategies. The median 30-year performance indicators for the evaluated management approaches across the 153 test locations are compiled in Table 9.



(a)



(b)

Figure11: LCCA Results (a) NPV distribution, (b) Cost vs condition Trade off

Table 9: 30 Years NPV worth of each strategy

Strategy	Median NPV ($/km)	Median Min PCI	% Time < 55	% Time < 40	Median N Treatments
Do Nothing	$230,587	25.7	58.1%	11.2%	0
ML-LCCA	$609,316	48.9	6.5%	1.1%	4
Fixed-Inspection	$620,693	60.2	0%	0%	3
Fixed 8-Year	$1,109,284	61.8	0%	0%	3

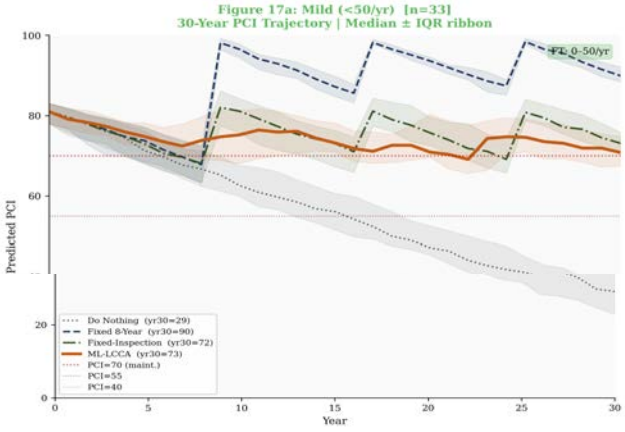
3.3.3 PCI Trajectory by FT Severity Band

Figure 12 presents the 30-year median PCI trajectory by FT severity band and maintenance strategy. The red dotted line at PCI = 70 marks the maintenance trigger threshold. The dark red line at PCI = 55 marks the thin overlay threshold, and the maroon line at PCI = 40 marks the mill and overlay threshold.

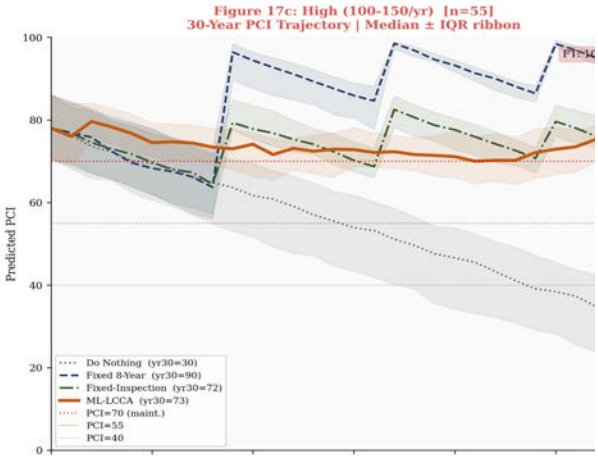
In the Mild FT band (Figure 14a), all three active strategies maintain median PCI above 60 throughout the 30-year horizon. ML-LCCA and Fixed-Inspection trajectories are nearly indistinguishable. This reflects the classifier's conservative triggering on sections that are deteriorating slowly. In the Moderate and High FT bands (Figures 14b and 14c), the three active strategies diverge progressively from Do Nothing after year 8. ML-LCCA maintains a median terminal PCI of approximately 55 to 60 at year 30, compared to below 30 for Do Nothing. The IQR ribbons are widest for ML-LCCA in all bands. This reflects that the classifier fires at different years for different sections, a direct consequence of section-level prediction that distinguishes ML-LCCA from Fixed 8-Year scheduling. In the Severe FT band (Figure 14d), the sample size is small ($n \leq 8$). ML-LCCA shows a variable trajectory reflecting section-level prediction variability. The limited sample size in this band suggests that the results should be interpreted with caution.

(a)

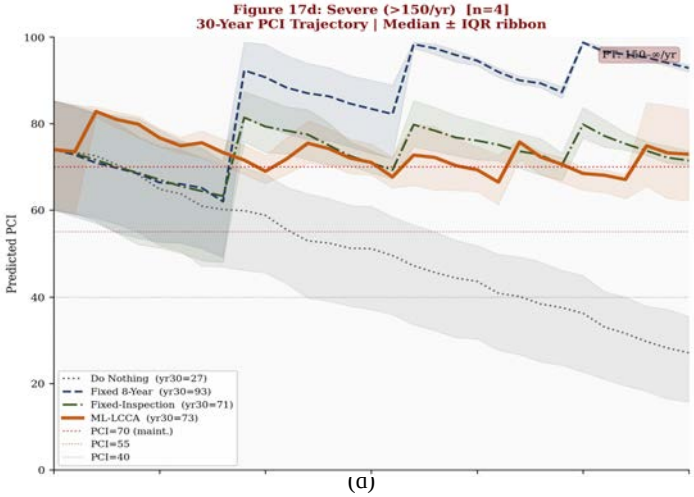
(a)



(b)



(c)



(d)

Figure 12: PCI Trajectory by FT Severity Band {a} Mild , (b) Moderate, (c)High severe section

4.0 Implications for Practice and Policy

The results support a shift from fixed-interval rehabilitation to condition-based scheduling in cold regions. Agencies can use the support vector classifier to flag sections at risk of falling below PCI 70 before the next inspection cycle. This predictive capability allows maintenance crews to focus on sections that truly need attention rather than following a calendar-based routine. The high recall of the model means that most deteriorating sections will be identified. The moderate false alarm rate is acceptable in practice because an unnecessary inspection costs far less than a missed deterioration event. Agencies can adjust the probability threshold upward if inspection budgets are tight or

downward if they want to be more conservative.

The feature importance analysis shows that detailed distress data adds value beyond composite indices. Total deduct value and alligator cracking deduct value are strong predictors of future maintenance needs. Agencies should continue to collect and code distress data according to ASTM standards. This information not only predicts failure but also guides treatment selection. Sections with high alligator cracking require structural overlays, while sections with transverse cracking may only need crack sealing. The interaction between freeze-thaw cycles and asphalt thickness suggests that structural design moderates climate vulnerability. Thin asphalt layers in high freeze-thaw areas deteriorate faster. Agencies should consider this interaction when designing new pavements. Increasing asphalt thickness in these zones may reduce long-term costs even if initial construction expenses are higher.

The group K-fold validation approach should become standard practice for any agency or vendor developing prediction models with LTPP data (Neves et al., 2022; Pereira et al., 2026). Standard cross validation overestimates performance because it ignores temporal correlation (Wang et al., 2023). Requiring group-based validation ensures that model accuracy reflects real-world conditions for new sections. The LCCA simulation demonstrates that machine learning triggers can reduce life cycle costs without sacrificing structural adequacy. The savings come from avoiding premature treatments on sections that remain adequate. Agencies can use this framework to evaluate maintenance strategies and justify investments in data collection and modelling capabilities.

5.0 Conclusion and Limitations

This study integrated machine learning with lifecycle cost analysis to support freeze-thaw pavement maintenance decisions. The classification model achieved a strong recall rate of 81.8% at the PCI 70 threshold, correctly identifying 89 of 109 sections that needed maintenance. The ASTM deduct values and structural interaction terms proved valuable for prediction. The LCCA simulation showed that the condition-responsive ML-LCCA strategy reduced median 30-year agency costs by 45.1% compared to the Fixed 8-Year schedule and by 1.8% compared to Fixed-Inspection, while maintaining pavement condition above critical thresholds.

Beyond this headline cost reduction, the study makes three further quantified contributions. First, section-level deterioration rates varied far more *within* individual freeze-thaw bands (standard deviation 1.78 to 3.02 points per year) than *between* band medians (approximately 1.0 point per year), demonstrating that band-average deterioration models are insufficient and validating the need for section-level prediction. Second, the selected classifier achieved a miss rate of 18.2%, substantially lower than four of the five alternative models evaluated, meaning fewer sections needing maintenance go undetected under the proposed framework than under comparable classifiers. Third, the Do Nothing baseline was shown to leave pavements below the critical PCI = 55 threshold for 58.1% of the 30-year simulation horizon, compared with only 6.5% under ML-LCCA, quantifying the structural-risk cost of deferred maintenance that fixed and reactive scheduling approaches carry.

Several limitations should be acknowledged. First, the dataset was limited to 153 sections with complete feature records. The severe freeze-thaw band had fewer than eight sections, which limits the generalisability of findings for that climate band. Second, the LCCA simulation used fixed treatment costs and assumed a constant deterioration rate. Actual costs vary by region and contractor, and deterioration rates change with climate conditions. Third, the simulation did not include user costs such as vehicle operating costs or delay costs during construction. Including these costs would increase the economic advantage of condition-based maintenance. Fourth, the probability threshold of 0.35 was selected based on the F2 score optimization. Different thresholds may be appropriate for agencies with different risk preferences. Fifth, the model was trained and tested on historical LTPP data. Performance may degrade when applied to sections with different design standards or traffic patterns.

Future research should extend this work in several directions. The model should be validated on independent datasets from other climate zones. The LCCA framework should be expanded to include user costs and environmental impacts, consistent with the broader lifecycle cost accounting principles established in the foundational LCCA literature (Walls & Smith, 1998). The classification framework should be extended to multi-class prediction that recommends specific treatment types. The SHAP analysis should be used to develop interpretable decision rules that agencies can implement without requiring machine learning expertise. Finally, the model should be embedded in a decision support system that provides real-time maintenance recommendations based on current condition data.

References

Abaza, K. A. (2024). A novel life-cycle analysis approach using pavement performance curves with cost and performance considerations. *International Journal of Pavement Engineering*.

https://doi.org/10.1080/19397038.2024.2405570

Alqaydi, S., Zeiada, W., Wakil, E., & Alnaqbi, A. (2024). Automated pavement condition index assessment with deep learning and image analysis: An end-to-end approach. *Sensors*, *24*(7), 2333. https://doi.org/10.3390/s24072333

Ali, A. A., Milad, A., Hussein, A., Yusoff, N. I. M., & Heneash, U. (2023). Predicting pavement condition index based on the utilization of machine learning techniques: A case study. *Journal of Road Engineering*, *3*, 266–278. https://doi.org/10.1016/j.jreng.2023.04.002

Applied Research Associates (ARA). (2015). *Ontario pavement life cycle cost analysis*. Applied Research Associates Inc.

ASTM International. (2016). *ASTM D6433-16: Standard practice for roads and parking lots pavement condition index surveys*. ASTM International.

Basnet, K. S., Shrestha, J. K., & Shrestha, R. (2023). Pavement performance model for road maintenance and repair planning: A review of predictive techniques. *Digital Transportation and Safety*, *2*(4), 253–267. https://doi.org/10.48130/DTS-2023-0021

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Dong, Q., & Huang, B. (2021). Investigation of factors contributing to pavement distress using machine learning. *Journal of Transportation Engineering, Part B: Pavements*, *147*(3), 04021024. https://doi.org/10.1061/JPEODX.0000283

Elkins, G. E., Thompson, T., & Gault, S. (2019). *Long-term pavement performance information management system pavement performance database user reference guide* (FHWA-HRT-18-082). Federal Highway Administration.

Federal Highway Administration (FHWA). (2001). *Long-term pavement performance studies: Pavement performance data user reference guide* (FHWA-RD-01-113). Federal Highway Administration.

Federal Highway Administration (FHWA). (2024). *InfoPave: LTPP data analysis portal*. https://infopave.fhwa.dot.gov

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Ghodratabadi, M., Golroo, A., & Entezari, M. S. (2025). Machine learning for predicting pavement roughness and optimising maintenance. *Road Materials and Pavement Design*. https://doi.org/10.1080/14680629.2025.2454376

Guo, M., Kovalskiy, V. P., Nian, T., & Li, P. (2024). Low-temperature cracking and improvement methods for asphalt pavement in cold regions: A review. *Buildings*, *14*(12), 3802. https://doi.org/10.3390/buildings14123802

Henderson, V., Tighe, S. L., & Norris, J. (2006). *LTPP data analysis: Pavement performance in northern climates* (FHWA-HRT-06-121). Federal Highway Administration.

Jiang, Y., Ullah, S., Zapata, C., & Yu, X. (2025). Assessment of the impacts of climate change on cold region pavement design and operation. *Journal of Infrastructure Preservation and Resilience*, *6*(1), 1–15. https://doi.org/10.1186/s43065-025-00126-2

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*. https://arxiv.org/abs/1412.6980

Li, X., Zhang, Y., & Chen, H. (2024). Improved model for pavement performance prediction based on recurrent neural network using LTPP database. *International Journal of Transportation Science and Technology*. https://doi.org/10.1016/j.ijtst.2024.07.001

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281–297. University of California Press.

Ministry of Transportation Ontario (MTO). (2019). *Ontario's transportation technology action plan — pavement management standard*. Ministry of Transportation Ontario.

Neves, J., Fonseca, P., & Antunes, A. (2022). Machine learning for pavement performance: Influence of cross-validation strategy on generalisation accuracy. *Road Materials and Pavement Design*, *24*(3), 756–773. https://doi.org/10.1080/14680629.2022.2032423

Pereira, T., Tran, N., Gadhoumi, K., & Pham, D. (2026). Cross-validation under data dependence: A review-derived taxonomy and trust, a leakage-aware Python toolkit. https://doi.org/10.13140/RG.2.2.18390.49929

Piryonesi, S. M., & El-Diraby, T. E. (2020). Data analytics in asset management: Cost-effective prediction of the pavement condition index. *Journal of Infrastructure Systems*, *26*(1), 04019036. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000512

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*, 6638–6648.

Santos, J., & Ferreira, A. (2013). Life-cycle cost analysis system for pavement management at project level. *International Journal of Pavement Engineering*, *14*(1), 71–84. https://doi.org/10.1080/10298436.2011.618535

Tamagusko, T., & Ferreira, A. (2023). Machine learning for prediction of the international roughness index on flexible pavements: A review, challenges, and future directions. *Infrastructures*, *8*(12), 170. https://doi.org/10.3390/infrastructures8120170

Tighe, S. L. (2001). Guidelines for probabilistic pavement life cycle cost analysis. *Transportation Research Record*, *1769*(1), 28–38. https://doi.org/10.3141/1769-04

Van der Weijde, A., Ouyang, J., & Overbeke, T. (2025). Optimal condition-based maintenance of asphalt-concrete pavement systems. *Road Materials and Pavement Design*. https://doi.org/10.1080/14680629.2025.2501713

walls, J., & Smith, M. R. (1998). *Life-cycle cost analysis in pavement design — in search of better investment decisions* (FHWA-SA-98-079). Federal Highway Administration.

Wang, H., Li, M., & Ozbay, K. (2023). Group-aware machine learning for pavement performance prediction using LTPP data. *Transportation Research Record*, *2677*(5), 1283–1295. https://doi.org/10.1177/03611981231156063