

The Pavement Condition Index Prediction Method Based on the PSO-SVM Model

Wenyuan Xu

Northeast Forestry University

Zehao Yang

Northeast Forestry University

Yongcheng Ji

yongcheng.ji@126.com

Northeast Forestry University

Ping Huang

China Railway (Heilongjiang) Expressway Investment Co., Ltd.

Article

Keywords: Asphalt Pavement Performance Prediction, Support Vector Machine, Particle Swarm Optimization, BP Neural Network, Random Forest

Posted Date: December 4th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-8228302/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

The Pavement Condition Index Prediction Method Based on the PSO-SVM Model

Wenyuan Xu¹, Zehao Yang¹, Yongcheng Ji^{1*}, Ping Huang²

Abstract: With more highways built in China and more Chinese traveling on them, asphalt pavement is becoming increasingly difficult to repair and maintain. To address the pavement performance prediction problems in this essay, we will present prediction methods for PCI using SVM, BPNN, and PSO-SVM. Then, all these influencing factors are analyzed using a random forest model to determine their importance. Using field data from a portion of an ordinary highway and a section of an expressway, this investigation develops a pavement performance prediction model. It examines how factors such as road age, AADT, average annual temperature, annual precipitation, and relative humidity affect PCI values. From the results, we can see that the PSO-SVM model is accurate and stable for nonlinear, high-dimensional data and has strong generalization performance. As shown by the Random Forest, PCI is influenced by factors such as road age, traffic, and temperature. This will help maintain the roads' pavement. It provides sound equipment for highway preservation administration, enabling better use and greater longevity on roads.

Keywords: Asphalt Pavement Performance Prediction, Support Vector Machine, Particle Swarm Optimization, BP Neural Network, Random Forest

¹Northeast Forestry University, China. ²China Railway (Heilongjiang) Expressway Investment Co., Ltd. China. *email: yongchengji@126.com

1. Introduction

Since the Reform and Opening-Up, China's road network has developed rapidly. By the end of 2024, the total length of roads exceeded 5.4904 million kilometers, as shown in Fig. 1. It also states that the total length of expressways exceeded 190,700 kilometers[1]. In fact, it is essentially a network on a road—a service-oriented road network—that provides basic services for economic and social development. At the same time, there is an ongoing increase in the number of cars on the roads, as well as overload, and together with these environmental and climatic factors, they also increase the degradation of the road. With limited funds but ever-increasing management requirements, the contradiction becomes more serious. LTPP and many other long-term pavement performance studies show that PMS helps us understand how well pavements perform in the future. Thus, it is imperative to quickly evaluate pavement performance and make highly accurate forecasts to conduct careful management of road property and preempt road maintenance.

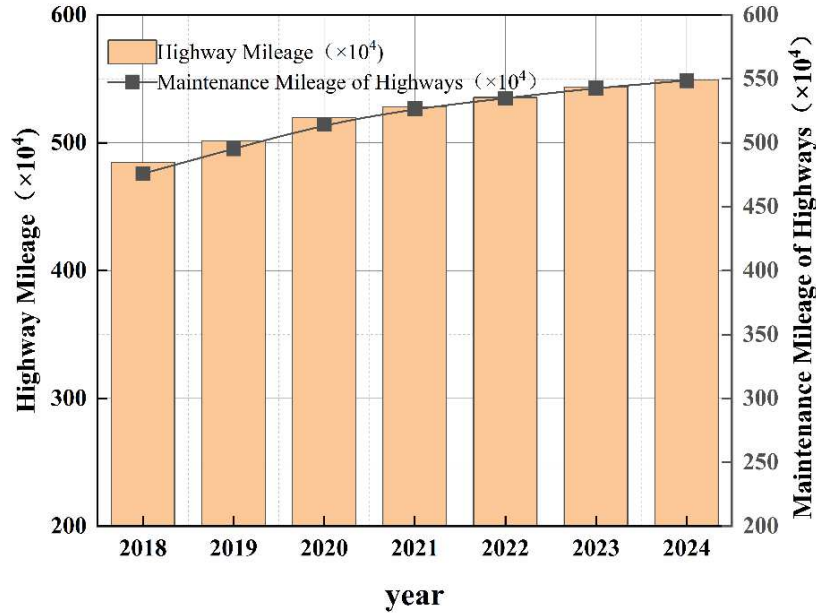


Figure 1. Total and Maintenance Road Mileage (2016-2022)

The types of performance prediction models in pavement management systems (PMS) are not highly varied. It mainly falls into three categories: deterministic, stochastic, or machine learning. Deterministic models produce a single pavement condition forecast value from a given set of conditions and timespan. In stochastic models, the randomness of pavement condition evolution is considered, and the pavement status probability distribution at any given time is generally provided, with the Markov model being the most representative. Due to advances in fields such as math and computing, various machine learning models have emerged, including SVMs, KNNs, and ANNs. Machine learning learns about correlation and structures in data; we can use it to reason and predict things that are not simple. They mostly involve many factors and are not linear. In deterministic models, Onayev and Swei[2] proposed an IRI degradation model for asphalt pavements that captures time-varying effects from construction year and maintenance improvements and emphasizes the importance of including explanatory variables, such as construction year, in network-level models. These models have the advantage of providing mechanistic explanations, but they face accuracy limitations when dealing with multi-source nonlinear coupling and regional transfer. In recent years, for network-level and multi-index prediction scenarios, a

growing body of work has combined temporal deep learning with automatic hyperparameter optimization. For example, Sun et al.[3] constructed a data matrix containing PCI, RDI, RQI, and SRI based on the Guizhou highway network and developed a multi-output LSTM prediction model using Bayesian optimization, achieving network-level prediction of four pavement performance indicators for the "next year." Gong et al.[4] utilized the large-scale LTPP database and introduced Random Forest Regression (RFR) to establish relationships between roughness and factors such as traffic, climate, and structure, proposing a new paradigm for predicting IRI using multi-source data. Yang et al.[5] proposed a machine learning-based framework for predicting pavement performance on highways in Xinjiang, integrating multiple machine learning models, including BP neural networks, PSO-BP neural networks, and Random Forests, demonstrating the superiority of PSO-BP neural networks in prediction accuracy.

For the joint prediction of multiple indicators of asphalt pavement, Xiao et al.[6] proposed a combined model that integrates Particle Swarm Optimization (PSO) and Back Propagation Neural Network (BPNN) to predict the functional and structural performance of asphalt pavements. The model optimizes the BPNN using PSO and employs marginal contribution to measure input influence, jointly modeling functional and structural indicators. It validates the effectiveness of coupling metaheuristics with learners. More directly related to this paper is the Support Vector Machine (SVM) approach: Yan et al.[7] applied Particle Swarm Optimization (PSO) to optimize both the penalty factor and kernel parameters of the SVM, using multiple pavement performance indicators as inputs to evaluate PCI, achieving results superior to those from experience-based parameter selection. Li et al.[8] applied the PSO-Support Vector Regression framework to years of highway inspection data, achieving faster convergence and lower error, thereby providing an effective baseline for small sample scenarios. Li et al.[9] proposed an improved Firefly Algorithm combined with SVM for highway pavement performance prediction, further enhancing stability and generalization. Wang et al.[10] applied Grey Relational Analysis for feature selection, coupled with Support Vector Regression, achieving robust predictions under small sample conditions.

In recent years, machine learning methods have been widely applied to predict asphalt pavement performance. For example, Sun et al.[11] developed an interpretable multi-output LSTM pavement performance degradation prediction model, combined with automatic hyperparameter optimization, achieving superior performance in network-level prediction compared to traditional baselines, while also providing feature contribution analysis to support maintenance decision-making. Zhao et al.[12] studied on the LTPP dataset and trained SVR, Random Forest Regression, GBM, and stacked ensemble models to predict IRI, rutting, cracking, and other distress indicators, while comparing their performance. Additionally, Li et al.[13] proposed the IFA-SVM method, which improves the Firefly Algorithm (IFA) to optimize the SVM hyperparameters (c, γ) globally. Compared with the unoptimized SVM and the traditional FA-SVM model, the prediction error and stability of highway pavement detection data can be significantly reduced. It can be observed as demonstrating the small-sample, high-nonlinearity scenario and as supporting maintenance decision-making. Guo et al.[14] built a Gradient Boosting Decision Tree (GBDT) model according to the LTPP database in order to predict IRI, rutting, and other indicators. By using grid search and cross-validation to optimize hyperparameters and combining sensitivity analysis to identify key factors, their results showed that GBDT outperformed both Random Forest and Artificial Neural Networks in terms of accuracy and robustness.

Tamagusko et al.[15] conducted a systematic review of machine learning studies on IRI prediction for flexible pavements and demonstrated that tree-based ensemble methods tend to be robust and accurate, serving as an advanced solution method in machine learning. Their benefit can help process a complex pavement performance dataset and its factors simultaneously with high accuracy. Gong et al.[16] presented the IRI prediction model, using Random Forest Regression, combined with traffic, climate, and structure multi-source factors. Compared with a traditional regression model, which significantly improves both accuracy and robustness, it can also be applied to network maintenance decisions. Mansour et al.[17] used field data from Louisiana to develop a long-term pavement performance (PCI) prediction model under a hot, humid climate. They also compared RF and CatBoost models, showing that the time horizon is up to 11 years. The results showed that Random Forest performed better than Catboost and can be applied to support asphalt pavement maintenance decisions. Ahmed et al.[18] compared the RF and XGBoost models for the prediction of SCI of pavements, they found that XGBoost has better prediction accuracy and stability than Random Forest (RF). Moghaddam et al.[19] came up with the SVM-FFA, a blend of the Firefly Algorithm and the Support Vector Machine, used to estimate the fatigue life of PET-modified asphalt mixtures. Compared to a single SVM model, SVM-FFA was more accurate and robust. Luo et al.[20] based on LTPP pavement data, used the MDI method of Random Forest to perform feature selection for IRI prediction. They then compared GBDT, XGBoost, SVM, and multiple least-squares methods. The ensemble model performed better, with GBDT performing the best. Wu[21] proposed a deep ensemble learning framework for multi-indicator pavement performance prediction. The approach combines the strengths of deep learning and tree-based ensembles to produce multi-output forecasts and enhances generalization and robustness through deep ensembling. Compared with single models or conventional ensembles, it delivers higher overall accuracy and is suitable for engineering-level maintenance decision support. Collectively, these studies indicate that models integrating deep learning and ensemble methods can effectively improve the accuracy and stability of pavement performance prediction, particularly when handling complex, high-dimensional data. However, most existing work focuses on single indicators and lacks comprehensive multi-indicator analysis. Future research should therefore consider hybrid schemes to jointly predict multiple asphalt pavement indicators.

Meanwhile, interpretable machine learning and uncertainty quantification have emerged as another line of work. Sandamal et al.[22] adopted an interpretable supervised framework to predict IRI on Sri Lankan trunk roads and used SHAP analysis to identify key drivers, improving model interpretability and transferability. Lv et al.[23] proposed an interpretable XGBoost approach that builds a high-quality input set via voting-based feature selection, enabling transparent identification of critical factors while maintaining accuracy. Taken together, these advances indicate that Support Vector Machines exhibit strong generalization in small-sample, highly nonlinear settings, with performance sensitive to the penalty parameter c and the kernel parameter γ [24]. Against this background, applying Particle Swarm Optimization (PSO) to global optimization of SVM hyperparameters can markedly improve prediction accuracy and stability without increasing data volume—this is the core motivation and contribution of the PSO-SVM method for PCI prediction proposed in this study[25].

2. Evaluation of Asphalt Pavement Performance

Because asphalt pavements are continuously impacted by vehicular loads and weather factors such as temperature, rain, snow, and sunshine, they are highly

susceptible to cracking, rutting, potholes, and other forms of distress, all of which degrade performance. These sorts of continuous, ongoing distress make driving uncomfortable, forcing us to lower our operating speed, posing safety risks, and adding to vehicle operation and road maintenance costs. Pavement will cease to provide service without prompt resolution. Therefore, it is necessary to regularly assess pavement conditions, take appropriate maintenance measures, and allocate resources accordingly to extend the service life of asphalt pavements. Management departments can quickly access real-time road surface information and select suitable repair and maintenance methods for weak parts. This is also one of the important purposes of asphalt pavement evaluation.

Pavement performance assessment indicators are generally divided into two categories: single indicators and composite indicators. An individual evaluation assesses specific aspects of performance, such as pavement condition, based on the type, degree, and distribution of distress. As for composite evaluation, which compares and summarizes several results of different parts into an overall assessment of the pavement's service level. Another way of saying this is that a single evaluation gives us just one point of view, while a composite evaluation covers the whole. According to current maintenance standards in China, the commonly used individual indicators are pavement condition index (PCI), ride quality index (RQI), structural strength index (SSI), side force coefficient (SFC) or pendulum friction number (BPN). The corresponding relationship is: the load-bearing capacity is represented by SSI; RQI represents the smoothness; SFC or BPN represents the slip resistance; and PCI represents the visible distress. The entire indicator system and process flow are shown in Fig. 2.

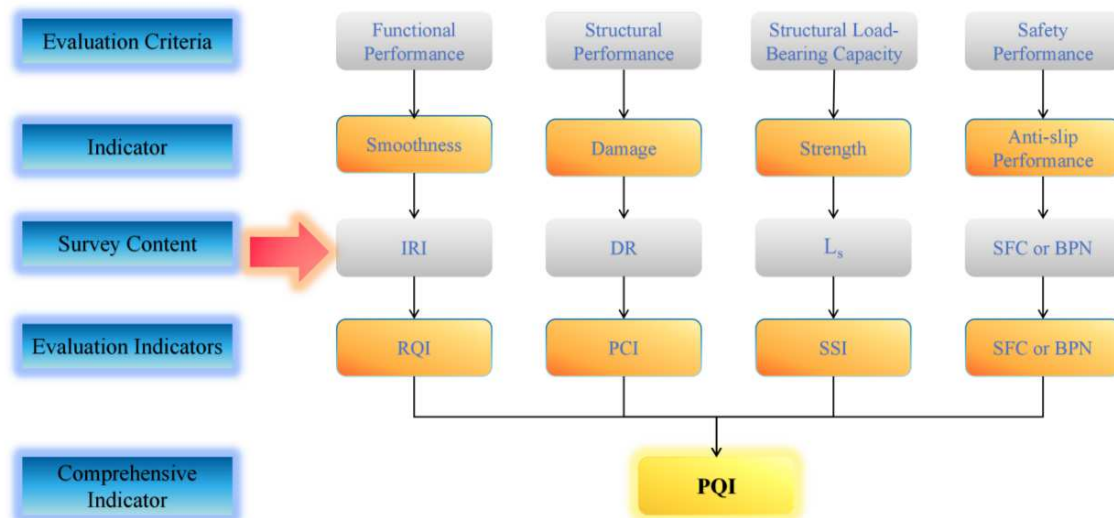


Figure 2. Pavement Performance Evaluation System

As per the "Highway Technical Condition Evaluation Standard"[26], the assessment of asphalt pavements comprises five indicators: pavement distress, rutting, smoothness, pavement structural strength, and pavement skid resistance (Figure 3). Every indicator is normalized to 0-100 from raw inspection data using standard methods, and better conditions are associated with higher scores. The individual results are then combined through weighing to form PQI.

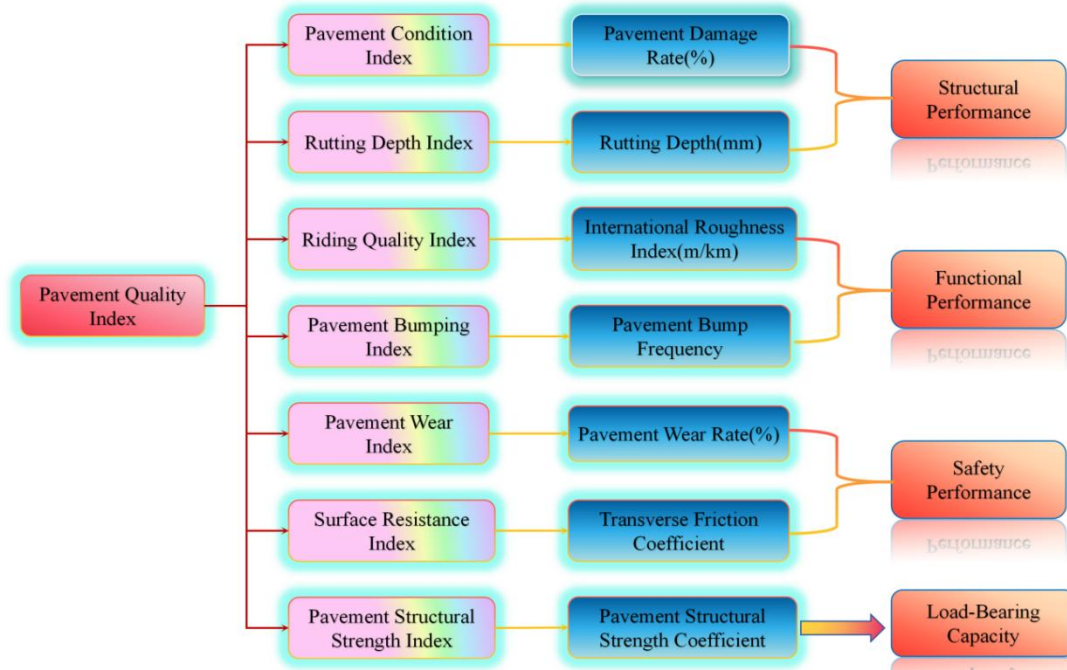


Figure 3. Highway Technical Condition Indicators Diagram

Though PCI fails to capture specific types of distress and damage patterns in asphalt paving paths, it can still serve as a comprehensive index to indicate the overall level of road surface damage. Therefore, this paper uses PCI as an important basis for preventive maintenance decision-making. Its calculation method and rating standards comply with the "Highway Technical Condition Evaluation Standard."

$$PCI = 100 - a_0 DR^{a_1} \quad (1)$$

$$DR = 100 \times \frac{\sum_{i=1}^{i_0} w_i A_i}{A} \quad (2)$$

Where:

DR —Pavement overall damage rate; A —Area of the surveyed pavement; A_i —Area of distress type i on the pavement (m^2); w_i —Weight or conversion factor for the i -th type of asphalt pavement distress; a_0 —Calibration coefficient, 15.00 for asphalt pavement; a_1 —Calibration coefficient, 0.412 for asphalt pavement; i —Distress type on the pavement; i_0 —Total number of distress types, 21 for asphalt pavement.

Table 1 shows the specific evaluation criteria for PCI as stipulated in the "Highway Technical Condition Evaluation Standard." The classification of these evaluation levels differs from the standard classification in the regulation.

Table 1. Pavement Distress Condition Evaluation Standards

Evaluation Indicators	Grade				
	Excellent	Good	Fair	Poor	Bad
Pavement Condition Index	≥ 90	[80,90)	[70,80)	[60,70)	< 60

To scientifically assess the treatment types and the applicability of preventive maintenance for sections of ordinary trunk roads requiring maintenance, industry technical standards have provided operational evaluation criteria and initiation conditions, as shown in Table 2.

Table 2. Preventive Maintenance Standards for Asphalt Pavements of Ordinary Trunk Roads[27]

Evaluation Indicators	Standard Name	Road Grade			
		First-Class	Second-Class	Third-Class	Fourth-Class
PCI	Highway Asphalt				
	Pavement Maintenance	≥80	≥75	≥75	≥70
	Technical Specifications				

The detection of asphalt pavement technical conditions is usually conducted through a combination of vehicle-mounted automation systems and manual ground inspection, with drone imagery used when necessary, as shown in Fig. 4. The vehicle-mounted multi-sensor system collects longitudinal profile data through inertial navigation and laser profile meters, which are then used to calculate the Ride Quality Index (RQI). Laser- or line-structured light sensors are used to measure the cross-sectional profile and assess rut depth (RDI). Meanwhile, high-definition cameras capture images for distress identification, providing data for calculating the Pavement Condition Index (PCI). For structural and safety performance, a falling weight deflectometer is commonly used to detect pavement deflection and calculate the structural modulus. At the same time, the Pendulum Friction Tester measures the Skid Resistance Index (SSI). Additionally, drone inspections can quickly capture high-resolution images without blocking lanes, identify surface distress such as cracks and loose material, and thus improve inspection efficiency and safety.



(a)



(b)

Figure 4. Pavement Inspection Equipment[28]

Data on asphalt pavements primarily come from pavement condition inspections, long-term monitoring, and traffic and climate data. Unlike bridges, the vast mileage and widespread distribution of asphalt pavements in the road network make it impossible to install long-term real-time monitoring systems over large areas. Therefore, most sections currently lack continuous sensor monitoring data. The common practice in China is for transportation authorities or inspection agencies to conduct regular road condition surveys every one or two years. The results are recorded in pavement technical condition evaluation forms or annual road condition reports and gradually entered into the Pavement Management System (PMS) database. These reports typically include the section location, pavement structure type, service life, distribution of major distresses, inspection methods, and corresponding performance evaluation indicators.

During the service life of the pavement, vehicle loads, construction methods, material properties, and environmental factors (such as temperature, rainfall, and freeze-thaw cycles) directly impact its performance evolution, leading to typical pavement distresses in asphalt surfaces, including crocodile cracking, rutting, transverse cracking, patching, longitudinal cracking, and strip patching. Pavement

inspection reports not only include the corresponding indicator values but also include photos of typical distresses, as shown in Fig5, providing a visual representation of the progression of pavement damage. This information serves as the most direct and comprehensive historical data for subsequent technical condition assessments, deterioration pattern analysis, and maintenance decision-making.

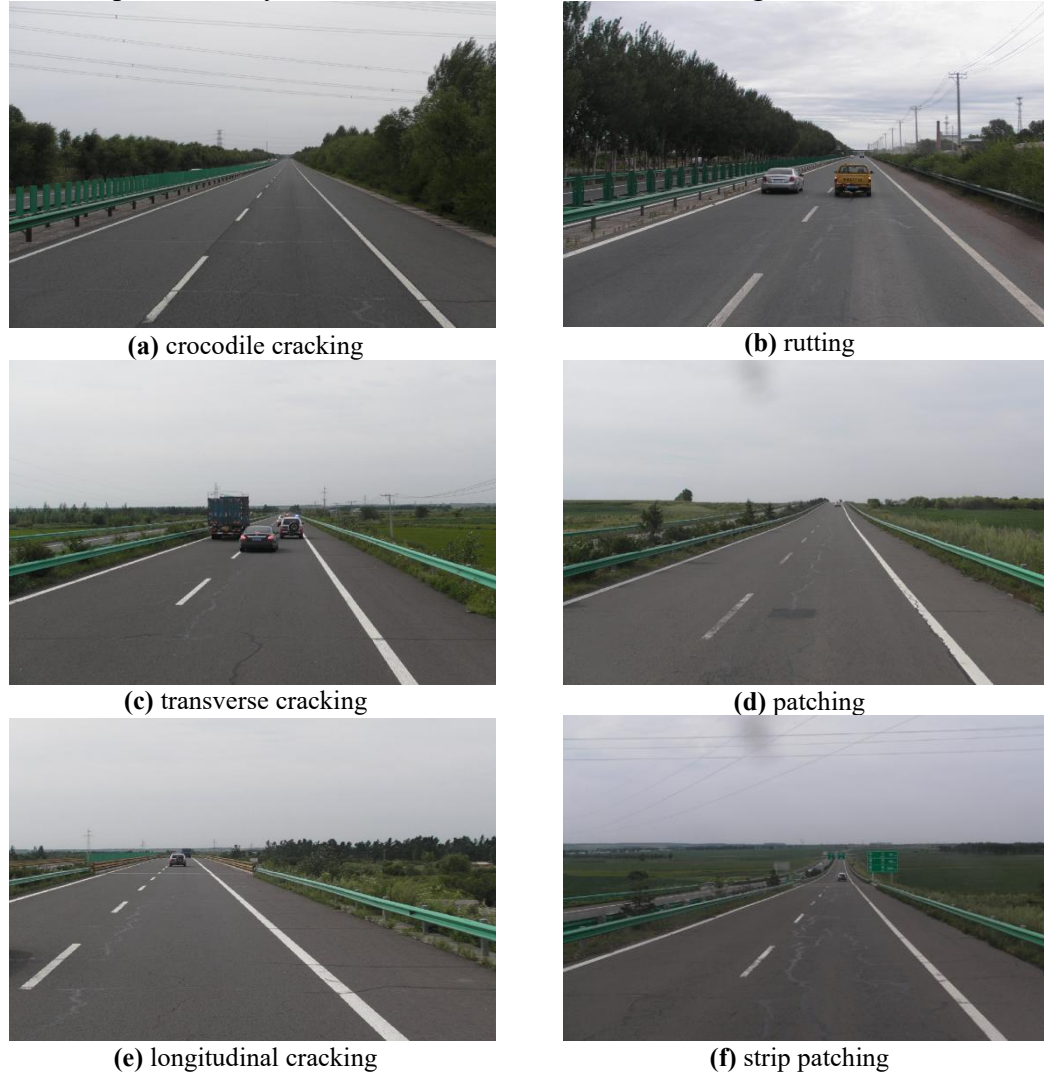


Figure 5. Pavement Distress Photos[28]

3. Data Processing

3.1 Statistical Analysis

The data used in this study mainly come from inspection data for a particular ordinary road (reference[29]) and a certain expressway (reference[30]), with PCI as the leading indicator. Additionally, road age, average annual daily traffic, average annual temperature, average annual precipitation, and relative humidity are considered. These indicators are not only reflections of the structural and functional state of pavement during the service period; they also reflect the combined effects of traffic load and environmental factors on pavement deterioration. Moreover, it will provide multidimensional support for the predictive models that follow this one. Tables 3 and 4 show the details.

Table 3. Expressways in a Certain Province[29]

Road Age /years	Annual Average Daily Traffic (AADT) / Vehicles	Annual Average Temperature / °C	Annual Precipitation / mm	Annual Relative Humidity / %	PCI
5	23568	12.38	514.35	59.5	98.22
6	25564	12.74	473.96	60.26	96.77
7	26806	13.07	463.88	57.31	95.82
8	28063	12.23	551.84	60.9	96.53
9	28934	11.03	633.55	67.08	93.29
10	29817	10.09	720.34	75.66	90.78
11	33133	10.8	681.2	37.49	90.9
12	32268	11.1	516.2	51.52	91.12
3	24265	12.38	514.35	59.5	100
4	26321	12.74	473.96	60.26	96.58
5	27598	13.07	463.88	57.31	97.08
6	28893	12.23	551.84	60.9	96.13
7	29790	11.03	633.55	67.08	94.92
8	30699	10.09	720.34	75.66	91.48
9	34113	10.8	681.2	37.49	94.28
10	32385	11.1	516.2	51.52	93.15
10	26814	7	205.3	36.44	94.71
11	29086	7.6	209.4	42.85	93.26
12	30498	7.5	185.2	44.48	93.26
13	31928	7.7	249.8	39.96	90.49
14	32920	6.8	184.5	38.14	91.19
15	33925	7.5	263.5	41.84	87.84
16	37697	7.4	182.4	35.79	89.76
17	34444	7.8	174.9	37.51	90.9
8	26838	8.1	74.3	32.14	84.94
9	29112	8.32	95.85	37.02	97.04
10	30525	9.61	142.65	37.61	96.31
11	31957	9.73	112.1	34.32	94.49
12	32950	8.53	117.45	35.87	94.62
13	33955	9.48	165.4	36.28	88.81
14	37731	8.79	54.4	32.58	88.44
15	34406	9.64	86.3	32.44	86.13

Table 4. Ordinary Roads in a Certain Province[30]

Road Age /years	Annual Average Daily Traffic (AADT) / Vehicles	Annual Average Temperature / °C	Annual Precipitation / mm	Annual Relative Humidity / %	PCI
1	8503	15.47	1233.26	90.64	95.59
2	7271	15.45	1261.07	92.64	94.63
3	8084	15.35	1143.39	90.66	92.11
4	5316	15.17	1176.09	92.51	87.76
5	4657	15.11	1258.58	91.99	83.21
1	7031	16.27	911.06	89.3	94.75
2	7875	15.4	1452.42	93.54	93.06
3	8342	15.86	1208.83	91.95	90.54
4	7407	15.94	1280.64	94	86.03
5	8261	15.87	1186.28	92.05	82.92
6	6979	15.73	1182.44	94.15	78.13
7	6951	15.59	1187.19	93.36	74.88
1	6812	16.06	1432.08	92.56	96.23
2	7543	16.4	1179.34	90.85	94.37
3	8194	16.54	1188.14	92.96	91.74
4	8587	16.46	1121.76	91.1	89.32
5	8881	16.29	1139.16	92.6	87.13
6	9328	16.2	1166.25	92.6	83.41
1	7112	16.39	1446.34	91.15	94.78
2	7579	16.66	1218.75	88.85	93.04
3	7598	16.86	1233.97	89.94	90.91
4	7256	16.78	1122.47	88.15	87.46
5	7011	16.63	1181.25	89.21	83.54
6	6886	16.51	1258.64	90.21	80.28
1	6452	18.27	1227.66	86.4	91.17
2	6754	19.3	1212	84.22	90.24
3	5791	19.11	1394.85	87.58	87.14
4	4484	19.09	1103.55	86.2	84.04
5	4796	18.43	1243.46	86.53	80.14
6	5027	18.35	1364.35	86.88	76.39

A total of 62 valid sample records were formed by combining data from both regions, with PCI values ranging from 74.88 to 100, road ages spanning from 1 to 17 years, and AADT ranging from 4,484 to 37,731 vehicles per day. The distribution of road age, annual average daily traffic (AADT), average annual temperature, annual precipitation, and annual relative humidity in relation to PCI is shown in Fig. 6. The overall trend indicates a negative correlation between PCI and road age. Under the same road age conditions, ordinary roads with higher traffic volumes show a faster decline in PCI, while expressways, benefiting from better maintenance conditions, exhibit greater stability. This difference highlights the combined impact of regional environment, traffic load, and road grade on pavement performance. It provides multi-

level data support for the subsequent development of predictive models based on SVM, BP neural networks, and PSO-SVM.

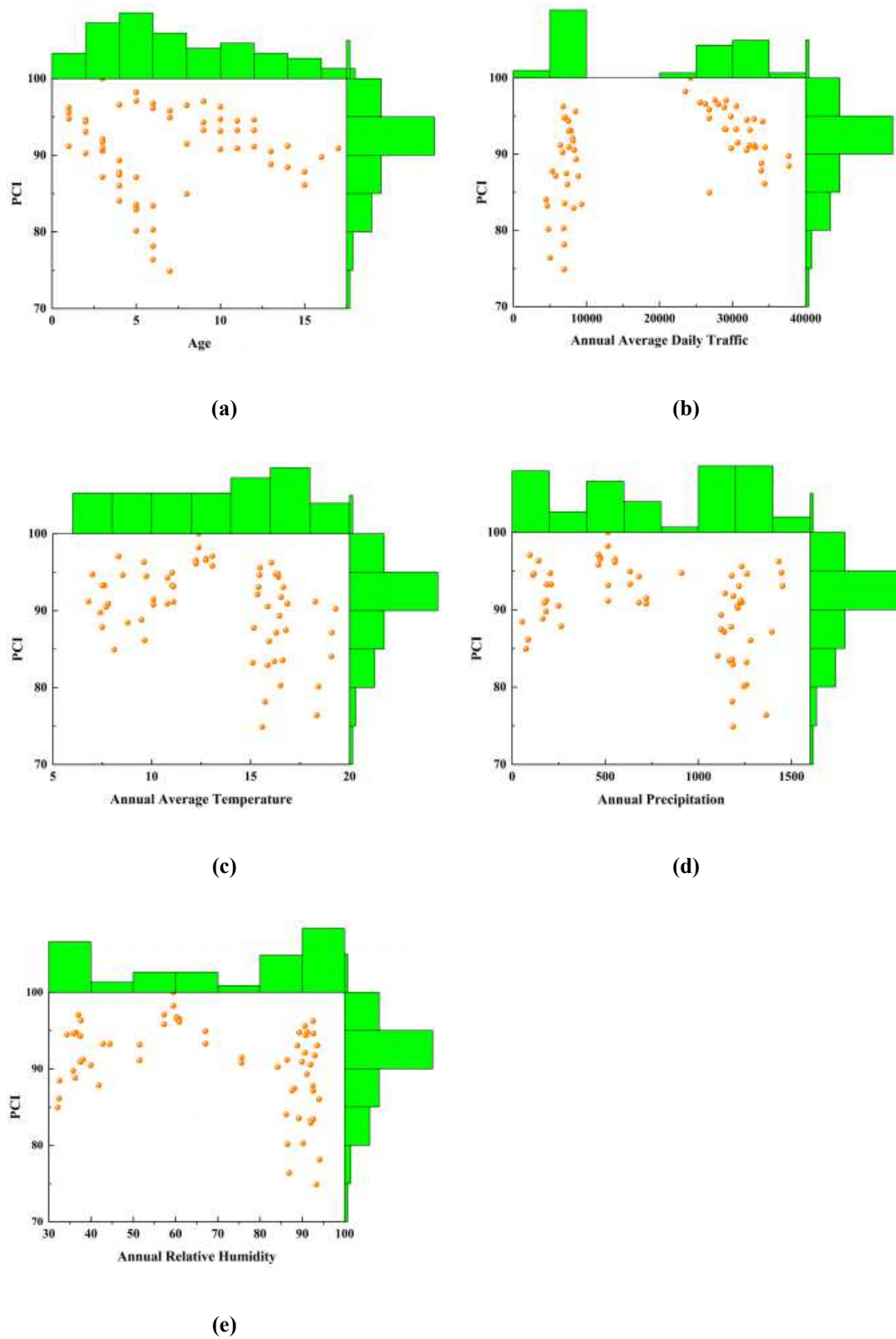


Figure 6. Data Distribution of PCI and Various Influencing Factors

3.2 Division of Training and Testing Sets

It is necessary to process the raw data systematically before training with machine learning models. Since there are significant differences in the ranges and units of different features, including them all directly in the model can lead to slower convergence and poorer predictive performance. Therefore, it is necessary to standardize all indicators to a standard scale, enabling comparisons of features and improving the reliability and accuracy of training models.

Therefore, we apply normalization to the dataset and linearly transform the original indicators to the $[0,1]$ range using a specified proportion. The data we get is processed so that data with different variable values maintain the ratio, and the influence of unit differences is omitted, ensuring the features have the same units and can be applied at the same level. The normalization formula is as follows:

$$y' = \frac{y - y_{\min}}{y_{\max} - y_{\min}} \quad (3)$$

Where y represents the original value of each evaluation indicator, $y_{\min}$ and $y_{\max}$ represent the minimum and maximum values of each evaluation indicator.

In model training, it is common to split the dataset into a training set and a test set, used for model learning and validation, respectively. The training set is used to reveal hidden patterns and features in the data and to improve the model's parameters; the test set is used to estimate the model's ability to predict new samples, thus evaluating the reliability of its predictions. For the 62 segment detection data from ordinary and expressway sections in a specific province, the data were randomly split at a 7:3 ratio, ensuring sample integrity. Among them, 44 samples comprised the training set, and 18 comprised the testing set. Such a division ensures there is sufficient data for the model's training and that a reasonable test can still be performed on the model's predictions using the independent test set, thereby providing the necessary data for building and validating the asphalt pavement PCI prediction model afterwards.

4. Simulation Testing

4.1 BP Neural Network

A BP (Back Propagation) neural network is a typical feedforward artificial neural network. It consists of an input layer, one or more hidden layers, and an output layer. It can be approximately and learning to solve complex problems through nonlinear methods. The PCI value of asphalt pavement can be affected by factors such as road age, road load, structural strength, climate, and pavement thickness in pavement performance prediction. In the past, traditional methods such as empirical formula methods or regression methods were often adopted but failed to capture the nonlinear relationships among these factors. Unlike a BP neural network with its multi-layer structure and weight update, it can automatically learn the complex relationship between input features and PCIP, enabling accurate predictions. However, BP neural networks rely on the error backpropagation algorithm for training, which can easily get stuck in local minima and fail to reach global optima. Additionally, when the network depth is too deep or the training sample size is large, extensive iterative calculations are required, leading to low training efficiency and potential overfitting, which hinder the model's generalization.

Based on the feature selection of factors influencing asphalt pavement performance, this paper selects five key indicators—road age, annual average daily

traffic (AADT), average annual temperature, annual precipitation, and annual relative humidity—as input variables for the BP neural network, and establishes a three-layer structured prediction model, as shown in Fig. 7. The input layer contains 5 neurons, and the output layer consists of 1 neuron, corresponding to the predicted value of the PCI. Based on the empirical formula: $h = \sqrt{n + m} + a$ (h is the number of hidden layer nodes, m is the number of input layer nodes, n is the number of output layer nodes, the value of a is generally an integer between 1 and 10), through multiple trial calculations and comparisons of accuracy and computational efficiency at different scales, the optimal number of hidden layer nodes was determined to be 6. During the training process, the following training parameters were used: a maximum of 1000 iterations, a target training error of 1×10^{-6} and a learning rate of 0.01.

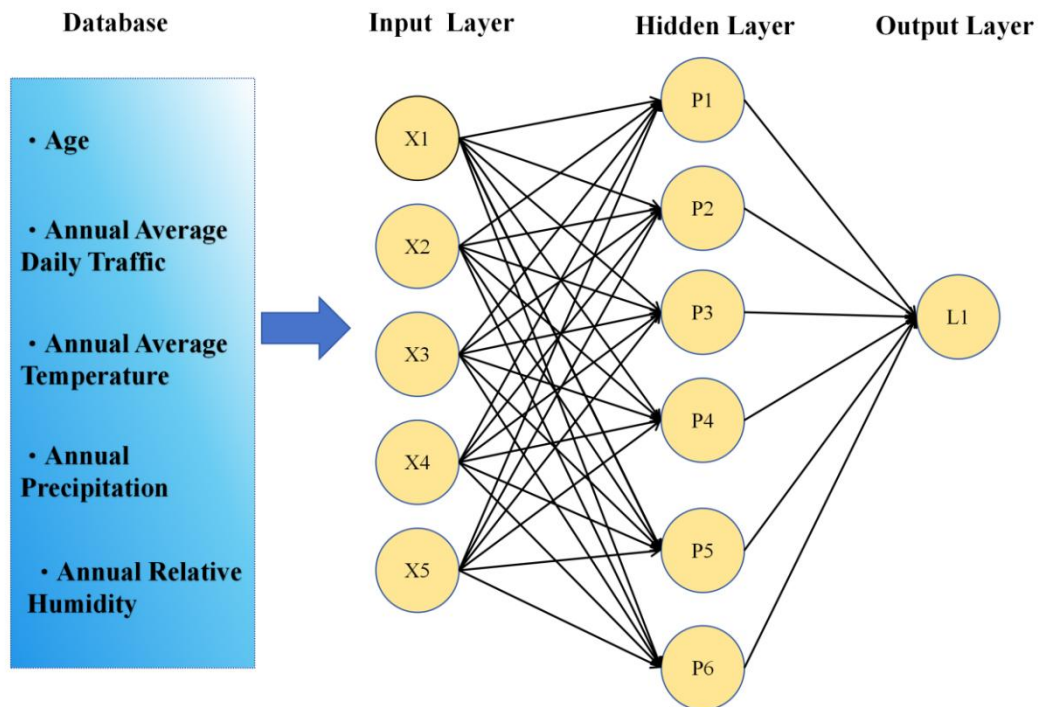


Figure 7. BP Neural Network Architecture Diagram

4.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a widely used machine learning algorithm for regression analysis. It is based on statistical learning theory[31] and can handle complex, nonlinear problems. The core idea of SVM is to use a kernel function to map the data into a higher-dimensional space, making originally linearly inseparable data separable in that space. Essentially, SVM seeks an optimal hyperplane that maximally separates samples from different categories, enabling classification or regression. For asphalt pavement performance prediction, notably the Pavement Condition Index (PCI), the SVM model can effectively capture the nonlinear relationships between pavement performance and various influencing factors (e.g., road age, traffic volume, climate). This makes it more advantageous in terms of prediction accuracy and applicability compared to traditional linear regression or empirical models.

Suppose SVM is used for asphalt pavement processing. In that case, it typically employs a kernel function to map the original data to a high-dimensional feature space, thereby overcoming the problem of linearly inseparable data. The linear, RBF, polynomial, and Sigmoid kernels are often used.

Due to the strong nonlinear characteristics of asphalt pavement data, the Radial Basis Function (RBF) kernel is the preferred choice in this study. The RBF kernel has

strong nonlinear mapping capabilities and requires fewer parameters, effectively handling complex multidimensional data and better capturing the patterns of pavement performance changes in higher-dimensional space. In this study, the same input and output features as the BP neural network were selected. The SVM classification model was constructed using Matlab programming software and its toolbox, with the model parameters set as follows: penalty factor $c=10.0$ and kernel function parameter $\gamma=0.1$.

4.3 Random Forest

Random Forest(RF) is an ensemble algorithm based on learning decision trees. It makes predictions and classifications by building several decision trees and combining their results. Compared to a single decision tree, Random Forest can significantly improve the model's prediction accuracy and stability by introducing randomness and diversity. In Random Forest, bootstrap sampling is used to randomly select subsets of the original data to create multiple training sets. At the same time, it also randomly selects some features for each decision tree split. To maintain the independence and diversity among trees to improve the generalization ability of the ensemble model.

To predict PCI values for asphalt pavements, we selected five input features related to pavement performance—road age, AADT, average annual temperature, annual precipitation, and annual relative humidity—with PCI value as the output variable. To avoid overfitting during model training, it limits the maximum number of decision trees in the Random Forest model to 50 and sets the minimum leaf size per tree to 1. This arrangement could balance the complexity of the model and the precision of prediction, and make the model more stable and better at generalizing.

4.4 Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) is a typical swarm intelligence evolutionary algorithm that simulates the process of group cooperation in nature, such as searching for food, to optimize complex functions. Unlike traditional gradient-based optimization methods, PSO does not rely on the differentiability of the objective function. Instead, it approximates the optimal solution through the interactions among multiple "particles" in the swarm. Each particle represents a candidate solution and has two attributes: position and velocity. During the iterative process, particles adjust their direction and movement magnitude based on their own historical best experience (personal best) and the best shared experience within the swarm (global best). This "dual memory mechanism" ensures a balance between global exploration and local exploitation. Meanwhile, the individual best and global best solutions are updated based on the fitness comparison of the current solution and the historical optimal solutions. The velocity update formula is as follows[32]:

$$v_{i,j} = \omega v_{i,j} + c_1 \cdot r_1 (p_{i,j} - x_{i,j}) + c_2 \cdot r_2 (g_j - x_{i,j}) \quad (4)$$

Where $v_{i,j}$ represents the velocity of particle i in the j -th dimension, ω is the inertia weight, $p_{i,j}$ represents the historical best position of particle i in the j -th dimension, g_j represents the value of the global best position in the j -th dimension, $x_{i,j}$ represents the position of particle i in the j -th dimension, c_1 and c_2 are the acceleration constants, also known as the learning factors. r_1 and r_2 are random numbers used to control the particle's exploration ability and its ability to exploit historical information.

Based on the updated velocity and position, the fitness function value for each particle can be recalculated to evaluate the quality of the current solution. The setup of the fitness function is based on the problem's purpose. For example, in predictive

modeling, we commonly use evaluation metrics such as MSE or RMSE. The algorithm compares the current particle's solution to its historical best solution's fitness and updates the individual best solution at every iteration. At the same time, update the ideal particle's position across the entire swarm to form a new global best solution.

4.5 Implementation Process of PSO-Optimized SVM

When using the SVM model, two hyperparameters— c and γ —affect its results. To make the SVM's parameters optimal through PSO to avoid overfitting or underfitting. PSO uses population-based search and iterative updates to find the best solution across a larger parameter space, improving the accuracy and speed of the SVM model's predictions. The basic flow chart of the algorithm is shown in Fig. 8 below:

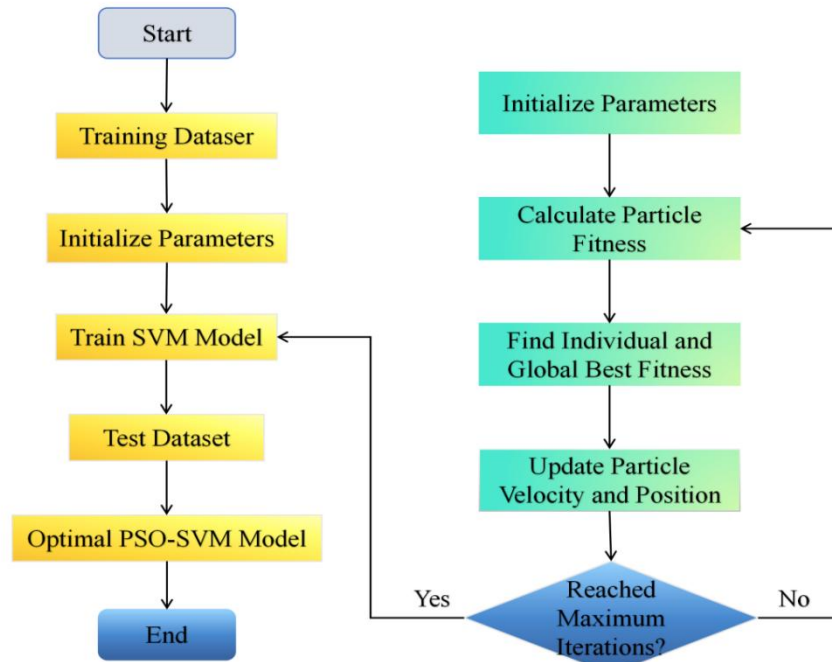


Figure 8. Flowchart of the PSO-SVM Prediction Model

The Particle Swarm Optimization (PSO) algorithm is used to optimize the Support Vector Machine (SVM) model. The specific details are as follows:

(1) Initialization: Set the particle swarm size, maximum number of iterations, learning factors, and other parameters. Randomly generate the initial positions and velocities of the particles.

(2) Fitness Calculation: Substitute the particle's corresponding parameters (c , γ) into the SVM model, calculate the prediction error as the fitness, and update the individual best and global best solutions.

(3) Update Position and Velocity: Adjust the particle's velocity and position based on the PSO update formula to generate new parameter combinations.

(4) Termination Condition: If the maximum number of iterations is reached or the error meets the accuracy requirements, stop; otherwise, continue iterating.

(5) Normalization: Normalize the input data to avoid the influence of dimensionality.

(6) Train the Optimal Model: Use the optimal parameters obtained from PSO to train the SVM model and perform prediction and validation.

In the testing process of the PSO-SVM model, the parameter settings are as shown in Table 5: the population size is 5, the maximum number of iterations is 100, c_1 and c_2 are set to 1.5 and 1.7, respectively. The search range for the penalty factor c and kernel parameter γ is set between 0.1 and 100. The fitness curve of the population

iteration is shown in the figure. As the iterations progress, the fitness value gradually converges and stabilizes. According to Fig. 9, after 54 iterations, the curve remains stable, indicating that the particle swarm has converged to the optimal solution. The optimal parameters obtained at this point are $c=1.48$ and $\gamma= 8.99$, which can be used for subsequent SVM model training and testing.

Table 5. PSO-SVM Parameters

Parameter	Population Size	Maximum Number of Iterations	c_1	c_2	Search Range for c and γ
Data	5	100	1.5	1.7	0.1-100

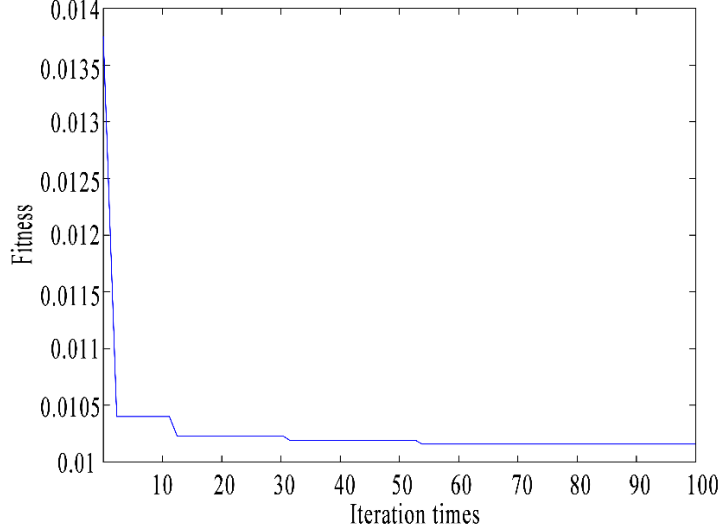


Figure 9. Fitness Curve Diagram

4.6 Model Evaluation Metrics

The coefficient of determination R^2 is used to measure the goodness of fit of a regression model. It represents the proportion of the total variation in the dependent variable that the independent variables explain. Its value ranges from 0 to 1, with values closer to 1 indicating better model fit and a higher proportion of explainable variation, thus leading to more ideal prediction results.

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

The Mean Absolute Error MAE is used to measure the average deviation of model predictions. It is calculated by averaging the absolute differences between the predicted and actual values for each sample, providing an intuitive error scale. Since MAE directly reflects the magnitude of the prediction bias, it is negatively correlated with prediction accuracy: the smaller the MAE , the lower the model's deviation from the target variable, indicating better overall prediction performance. Compared to squared-error metrics, MAE is less sensitive to outliers, providing a more objective measure of the average error across most samples.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

Root Mean Square Error converts the prediction bias into a value with the same units as the original data, making it easier to interpret in practical applications. Compared to MAE , $RMSE$ gives more weight to larger errors because it is calculated as the square root of the mean squared error, thereby highlighting the impact of

prediction points that deviate significantly from the true values on the overall error. The lower the RMSE value, the higher the model's prediction accuracy.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (7)$$

Where y_i , $\hat{y}_i$, and $\bar{y}_i$ represent the actual values, predicted values, and the mean of the actual values, respectively.

5. Model Evaluation and Results

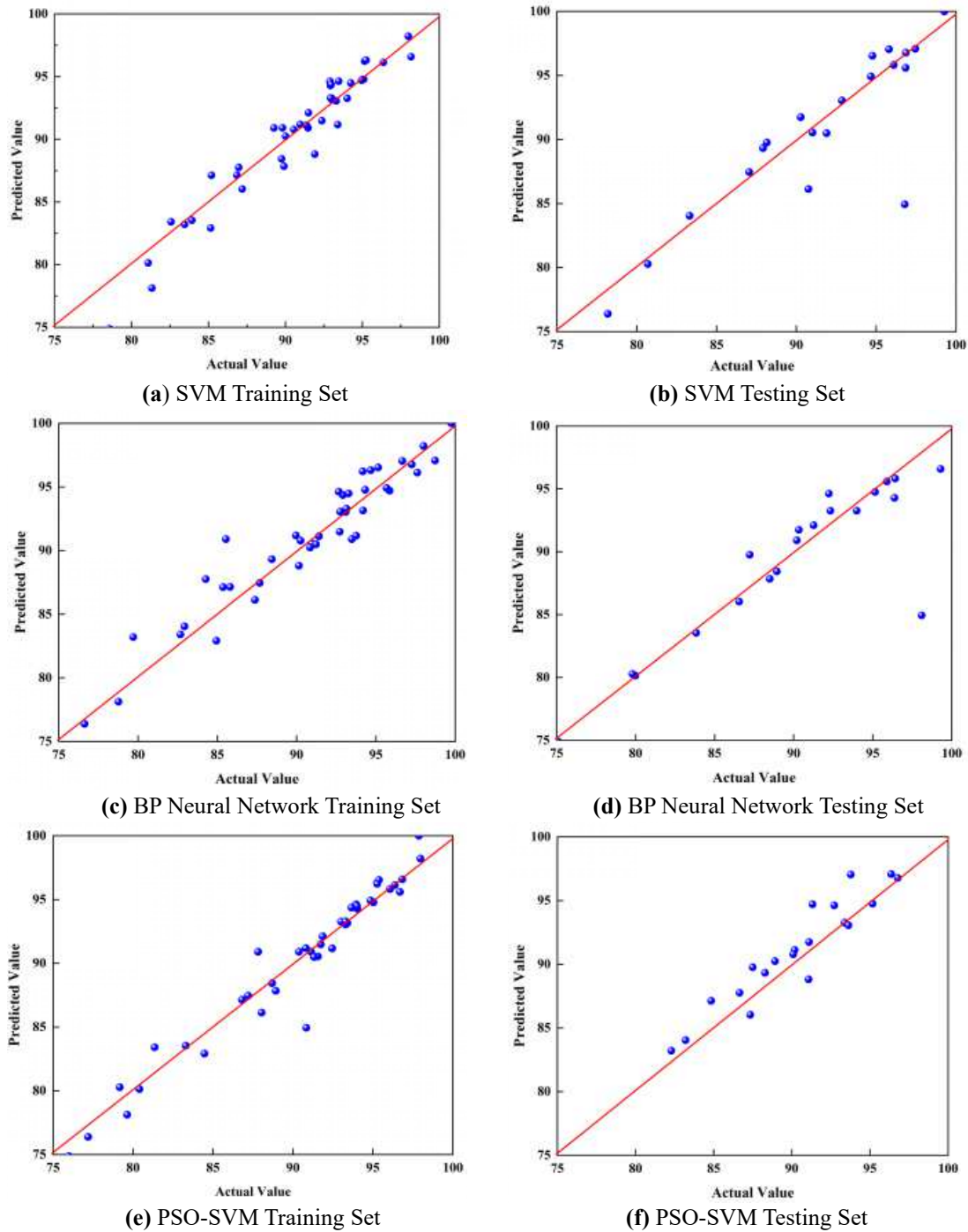


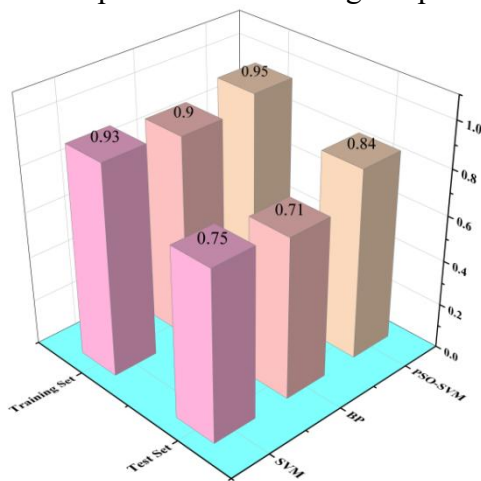
Figure 10. Model Training and Testing

To thoroughly evaluate the effectiveness of the PSO-optimized SVM model, this study systematically compares the predictive performance of the traditional SVM, BP

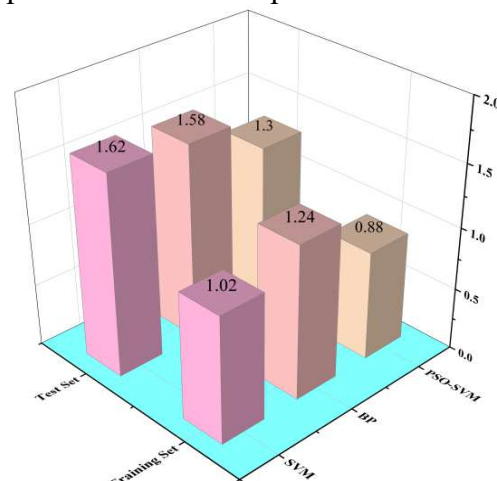
neural network, and PSO-SVM models on both the training and test sets. The specific training and testing results are shown in Fig. 10. Firstly, for the training set, the prediction results of all three models exhibit a noticeable linear trend. Both the SVM and PSO-SVM models show that their predicted values are closely aligned with the actual values along the diagonal, indicating that they are effective at capturing the central relationship between PCI and the input variables. The BP neural network also performs well on the training set, though it slightly overestimates in the lower score range (approximately 80-87). Overall, the error is small, and the model fits the training data quite well. In comparison, the PSO-SVM model's performance on the training set is similar to that of the SVM. However, its optimized hyperplane is more robust, enabling it to handle complex nonlinear relationships better and achieve better training performance.

However, on the test set, the SVM model's prediction accuracy drops relative to the training set; the scatter points are even further from the diagonal. Looking at the data, there is significant variance, and more interestingly, a single sample with an accurate value around 97 is way underestimated at 85, thus the model lacks strong generalization ability to unknown data. The BP neural network performs well on the test set, with overall predictions being relatively small and only a few points deviating from the reference line, indicating that it is still adapting to the test data. Compared with the SVM and BP neural network, the PSO-SVM model performs better on the test set, with test points more concentrated; the difference between the predicted and actual values is smaller, and the model's prediction accuracy improves significantly. This shows that the PSO-SVM model performs well on the training set and generalizes and stabilizes better when applied to new data.

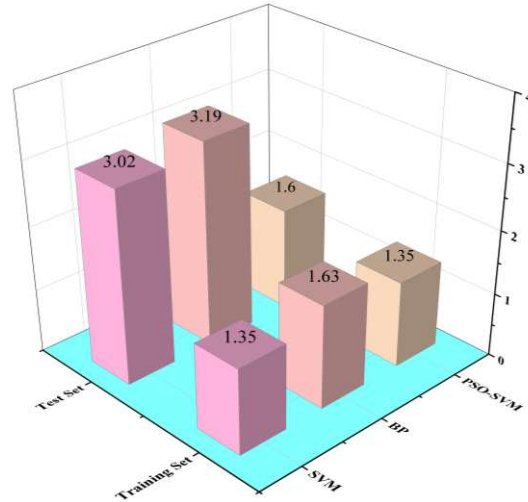
Overall, the PSO-SVM model is much more accurate and stable when processing large-scale data and nonlinear models than SVM and BPNN, especially in testing, where it performs far better than the others. While both BP Neural Networks and SVMs can already fit the training set well, their generalization performance on test sets is relatively weak. PSO-SVM, because it has an optimization capability, not only fits the training data better but also provides more accurate predictions for new data, suggesting that the optimized SVM has good predictive power for novel samples.



(a) R^2 Results



(b) MAE Results



(c) *RMSE* Results

Figure 11. Results of the Three Key Model Metrics

The evaluation criteria for the BP, SVM, and PSO-SVM models are shown in Fig. 11: on the training set, SVM achieves an R^2 of 0.93, but on the test set, it is only 0.75, indicating some overfitting. BP neural networks achieve an R^2 of 0.90, which is relatively good, but the drop on the test set is quite significant, with an R^2 of only 0.71, indicating poor generalization. The PSO-SVM model achieves the best results, with R^2 values of 0.95 in the training set and 0.84 in the test set, indicating good and strong fitting accuracy.

Regarding the Mean absolute error (*MAE*), the SVM model yielded errors of 1.02 and 1.62 on the training and test sets, respectively, indicating a decrease in test-set prediction accuracy. BP neural network is 1.24 on the training set and 1.58 on the test data, with a greater fluctuation. The PSO-SVM model has *MAEs* of 0.88 on the training set and 1.30 on the test set, which is the smallest error; therefore, it has the best generalization ability among the three models.

In terms of the *RMSE* (metric), the SVM model has an *RMSE* of 1.35 on the training dataset, but it rises to 3.02 on the test set, which implies that it is poorly fitting on the test set. BP neural network's *RMSE* is 1.63 on the train set and 3.19 on the test set, respectively, with significant error as well. PSO - SVM model has an *RMSE* of 1.35 on the training set and 1.6 on the testing set, with the smallest error and stable prediction.

On the whole, PSO-SVM performs best across the three metrics; the model also shows excellent stability and accuracy on the test set. So this model can be used to predict asphalt pavement performance in real-world settings with good generalization.

According to the importance analysis of influencing factors shown in Fig. 12, pavement condition (PCI) is affected by multiple factors. Specifically, the factors presented in the figure include X1 (road age), X2 (annual average daily traffic), X3 (average annual temperature), X4 (annual precipitation), and X5 (annual relative humidity). Based on the analysis of these factors' importance, pavement condition (PCI) is primarily influenced by road age, annual average daily traffic, and average annual temperature. Road age has the most significant impact on pavement condition, with an importance of 77.8%, clearly higher than the 49.8% impact of annual average daily traffic, a difference of 1.56 times. This suggests that as the service life of a road increases, the pavement gradually deteriorates, with aging sections requiring more maintenance.

While traffic volume also has a significant impact, it is smaller compared to road age, and its effect is more short-term. The influence of average annual temperature is 53.5%, slightly higher than that of traffic volume. Temperature fluctuations accelerate the aging of pavement materials, particularly in extreme climate conditions. Although the impact of temperature is smaller than road age, it has a lasting effect on pavement durability. Finally, it is seen that the effects of annual precipitation and relative humidity are even less on the pavement, and waterproofing and drainage precautions are needed in humid areas.

The conclusion is that the analysis results can also form the basis for future pavement maintenance strategies. Especially when funds are relatively short, identifying key influencing factors and allocating maintenance resources effectively can improve maintenance efficiency, extend the service life of roads, and reduce future maintenance costs.

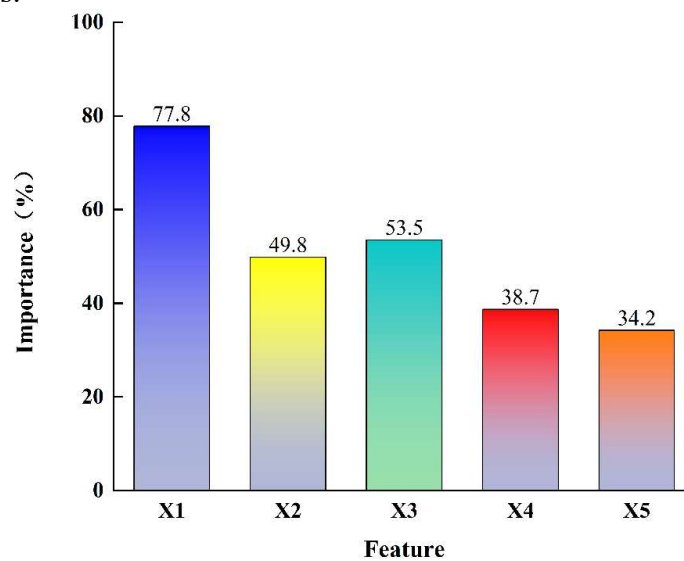


Figure 12. Random Forest Analysis of Factor Importance

6. Conclusion

According to the inspection data of asphalt pavement in ordinary and expressways of a province, this paper adopts SVM, BP neural network, and PSO-SVM (PSO-SVM) model to predict and analyze PCI. Also, a Random Forest will be applied to assess the importance of influencing factors. The main conclusions can be described as such:

(1) Data preprocessing, where the standardization method was used to standardize 5 input fields - road age, AADT, annual average temperature, annual precipitation, and annual relative humidity to be at the same scale for data, to prevent dimensional influence on model training. These processed features served as a reliable input basis for the training model.

(2) Compared to traditional SVM and BP neural networks, the PSO-SVM model showed excellent prediction results on the training set and testing set, and has specific generalization ability. PSO-SVM used the PSO algorithm to optimize SVM hyperparameters, eliminating the need for manual adjustment and thereby improving the SVM model's prediction accuracy. As shown in the test results, the PSO-SVM achieved the largest R^2 and the smallest MAE and $RMSE$, indicating strong predictive performance for PCI values of asphalt pavements.

(3) According to the result of the Random Forest model, the most important factors in the PCI prediction result were found to be the road age, AADT per lane, and average

annual temperature. Particularly, road sections with longer road age and higher traffic volumes typically exhibit poorer pavement conditions. Thus, when making judgments on highway management, one needs to prioritize these high-impact factors, allocate maintenance resources appropriately, and ensure stable, safe roads for the long term.

(4) The prediction result of this paper can give a scientific basis for the maintenance department of the highway. If we consider resource limits, the priority for upkeep funding must go first to road segments with high age, heavy traffic, and severe weather conditions, especially in zones with higher climate impact factors. Moreover, while annual precipitation and relative humidity contribute little to PCI, suitable drainage remains important in humid environments, as it delays the onset of pavement distress.

(5) The PSO-SVM model performance in this study shows that the model has some significant advantages that deal with complex nonlinear problems and makes it a suitable model to be used in predicting and managing asphalt pavement. As for future research, it would also be better to consider additional influencing factors, such as pavement structure types and climate changes across different seasons, to improve the model's predictive and generalizing capabilities. At the same time, combining other optimization algorithms with machine learning models could yield a more stable and accurate model, providing better support for Highway maintenance management.

Data availability

The datasets analyzed in this study are available from the corresponding author upon a reasonable request.

Funding

This work was supported by Research on the Development of a Technical System for the Evaluation of Road and Bridge Structural Conditions based on Damage Detection Results(HJK2023B009-5), Research and Application of Key Technologies for Smart Construction Site Monitoring and Inspection during Highway Construction(HJK2023B009) and Development Program of Heilongjiang (GZ2024009).

References

- [1] Ministry of Transport of the People's Republic of China. 2025. Statistical Bulletin on the Development of the Transport Industry, 2024. Beijing: Ministry of Transport. (in Chinese)
- [2] Onayev A, Sweil O. IRI deterioration model for asphalt concrete pavements: capturing performance improvements over time[J]. Construction and Building Materials, 2021, 271: 121768.
- [3] Sun X, Wang H, Mei S. Explainable highway performance degradation prediction model based on LSTM[J]. Advanced Engineering Informatics, 2024, 61: 102539.
- [4] Gong H, Sun Y, Shu X, et al. Use of random forests regression for predicting IRI of asphalt pavements[J]. Construction and Building Materials, 2018, 189: 890-897.
- [5] Yang Q, Tian W, Dai X. Machine Learning-Based Highway Pavement Performance Prediction in Xinjiang[J]. Infrastructures, 2025, 10(7): 189.
- [6] Xiao M, Luo R, Chen Y, et al. Prediction model of asphalt pavement functional and structural performance using PSO-BPNN algorithm[J]. Construction and Building Materials, 2023, 407: 133534.

- [7] Yan K Z, Zhang Z. Research in analysis of asphalt pavement performance evaluation based on PSO-SVM[J]. *Applied Mechanics and Materials*, 2011, 97: 203-207.
- [8] Li Z, Zhang J, Liu T, et al. Using PSO-SVR algorithm to predict asphalt pavement performance[J]. *Journal of Performance of Constructed Facilities*, 2021, 35(6): 04021094.
- [9] Li H, Lin M, Wang Q. Performance Prediction of Highway Asphalt Pavement Based on IFA-SVM[J]. *Journal of Highway and Transportation Research and Development (English Edition)*, 2020, 14(3): 20-27.
- [10] Wang X, Zhao J, Li Q, et al. A hybrid model for prediction in asphalt pavement performance based on support vector machine and grey relation analysis[J]. *Journal of advanced transportation*, 2020, 2020(1): 7534970.
- [11] Sun X, Wang H, Mei S. Explainable highway performance degradation prediction model based on LSTM[J]. *Advanced Engineering Informatics*, 2024, 61: 102539.
- [12] Zhao J, Wang H. Machine learning based pavement performance prediction for data-driven decision of asphalt pavement overlay[J]. *Structure and infrastructure engineering*, 2025, 21(6): 940-955.
- [13] Li H, Lin M, Wang Q. Performance Prediction of Highway Asphalt Pavement Based on IFA-SVM[J]. *Journal of Highway and Transportation Research and Development (English Edition)*, 2020, 14(3): 20-27.
- [14] Guo R, Fu D, Sollazzo G. An ensemble learning model for asphalt pavement performance prediction based on gradient boosting decision tree[J]. *International Journal of Pavement Engineering*, 2022, 23(10): 3633-3646.
- [15] Tamagusko T, Ferreira A. Machine learning for prediction of the international roughness index on flexible pavements: A review, challenges, and future directions[J]. *Infrastructures*, 2023, 8(12): 170.
- [16] Gong H, Sun Y, Shu X, et al. Use of random forests regression for predicting IRI of asphalt pavements[J]. *Construction and Building Materials*, 2018, 189: 890-897.
- [17] Mansour E, Mousa M R, Dhasmana H, et al. Machine-learning-based framework for prediction of the long-term field performance of asphalt concrete overlays in a hot and humid climate[J]. *Transportation Research Record*, 2023, 2677(10): 375-385.
- [18] Ahmed N S. Machine Learning Models for Pavement Structural Condition Prediction: A Comparative Study of Random Forest (RF) and eXtreme Gradient Boosting (XGBoost)[J]. *Open Journal of Civil Engineering*, 2024, 14(4): 570-586.
- [19] Moghaddam T B, Soltani M, Shahraki H S, et al. The use of SVM-FFA in estimating fatigue life of polyethylene terephthalate modified asphalt mixtures[J]. *Measurement*, 2016, 90: 526-533.
- [20] Luo Z, Wang H, Li S. Prediction of international roughness index based on stacking fusion model[J]. *Sustainability*, 2022, 14(12): 6949.
- [21] Wu Y. From ensemble learning to deep ensemble learning: A case study on multi-indicator prediction of pavement performance[J]. *Applied Soft Computing*, 2024, 166: 112188.

- [22] Sandamal K, Shashiprabha S, Muttill N, et al. Pavement roughness prediction using explainable and supervised machine learning technique for long-term performance[J]. Sustainability, 2023, 15(12): 9617.
- [23] Lv B, Gong H, Dong B, et al. An Explainable XGBoost Model for International Roughness Index Prediction and Key Factor Identification[J]. Applied Sciences, 2025, 15(4): 1893.
- [24] Wang X, Zhao J, Li Q, et al. A hybrid model for prediction in asphalt pavement performance based on support vector machine and grey relation analysis[J]. Journal of advanced transportation, 2020, 2020(1): 7534970.
- [25] Li Z, Zhang J, Liu T, et al. Using PSO-SVR algorithm to predict asphalt pavement performance[J]. Journal of Performance of Constructed Facilities, 2021, 35(6): 04021094.
- [26] Ministry of Transport of the People's Republic of China. Highway technical condition evaluation standard (JTG 5210–2018). China Communications Press, Beijing, China (2019). [In Chinese]
- [27] Lin, X. & Li, Q. Technical system for highway pavement maintenance decision-making. Highway 66(10), 357–364 (2021). [in Chinese]
- [28] Heilongjiang Provincial Highway Bureau. (2016). 2016 Heilongjiang Province high-grade highway pavement technical condition inspection and maintenance analysis report. National Engineering Research Center for Highway Maintenance Technology.
- [29] Chen, Chang-sheng. "Study on the Combination Model of Asphalt Pavement Performance Degradation for Ordinary Highways." Guizhou University, 2021. <https://doi.org/10.27047/d.cnki.ggudu.2021.002932>.
- [30] Li, H., Zhou, S., Li, Q., Liu, Z., & Jia, W. (2025). Condition index prediction of asphalt concrete pavement surface damage based on variable weight combination. Journal of Lanzhou University (Natural Sciences), 61(1), 35–42. <https://doi.org/10.13885/j.issn.0455-2059.2025.01.005>
- [31] Vapnik V N. An overview of statistical learning theory[J]. IEEE transactions on neural networks, 1999, 10(5): 988-999.
- [32] Tang K, Meng C. Particle swarm optimization algorithm using velocity pausing and adaptive strategy[J]. Symmetry, 2024, 16(6): 661.