geotechnics

MDPI

Article

Data-Driven Evaluation of Bearing Capacity for In-Service Pile Foundations Using Dynamic Stiffness and Machine Learning

Yuxuan Zeng [1], Jun Guo [2], Wangyu He [3], Yueying Chen [1] and Meng Ma [1,*]

[1] School of Civil Engineering, Beijing Jiaotong University, Beijing 100044, China; 24231173@bjtu.edu.cn (Y.Z.); 25120894@bjtu.edu.cn (Y.C.)
[2] School of Cyberspace Science and Technology, Beijing Jiaotong University, Beijing 100044, China; 24231176@bjtu.edu.cn
[3] School of Computer Science & Technology, Beijing Jiaotong University, Beijing 100044, China; 24231178@bjtu.edu.cn
* Correspondence: mameng@bjtu.edu.cn

Abstract

In the assessment of bearing capacity for in-service bridge pile foundations, static load tests are costly, destructive, and difficult to scale. The traditional dynamic formula approach relies heavily on an empirical dynamic–static conversion coefficient that introduces considerable uncertainty. To address these limitations, this study proposes a non-destructive evaluation method for pile foundation bearing capacity based on measured dynamic stiffness and machine learning algorithms. Using data from a highway bridge inspection project, a dataset comprising 680 piles was compiled, including measured dynamic stiffness, geometric parameters, and design load information. An end-to-end binary classification model was constructed to map multidimensional physical features to an engineering decision target, namely, whether the bearing capacity meets the design requirement. The performance of several algorithms was compared, including logistic regression, random forest, and gradient boosting decision tree (GBDT). Among the evaluated models, the GBDT model demonstrated the best capability for capturing the complex nonlinear pile–soil interactions. On an independent test set, it achieved an accuracy of 96.3% and an F1 score of 0.96, with a very low false-negative rate, satisfying the high precision required for engineering safety screening. Feature importance analysis indicates that measured dynamic stiffness contributed approximately 42% to the classification outcome, establishing it as the dominant indicator for detecting capacity deficiencies and reinforcing its physical relevance as a key health indicator for pile foundations. This study demonstrates that data-driven methods can effectively circumvent the uncertainty associated with traditional empirical coefficients, providing a promising approach to the health monitoring and rapid evaluation of in-service bridge pile foundations.

Keywords: pile foundation bearing capacity; dynamic stiffness; machine learning; non-destructive testing; feature importance

Academic Editor: Tadahiro Kishida

Received: 29 March 2026
Revised: 30 April 2026
Accepted: 12 May 2026
Published: 18 May 2026

1. Introduction

Pile foundations, as critical load-bearing components in large-scale infrastructure such as bridges, high-rise buildings, and port terminals, play a vital role in ensuring the safety and durability of the structure. With the continuous growth of global transportation infrastructure stock, a large number of bridges constructed in earlier periods have entered a critical stage of maintenance and rehabilitation. In particular, under heavy-haul railway conditions [1] and high-load highway traffic environments, pile foundations may experience performance degradation due to long-term dynamic loading and geological environmental effects. Therefore, regular inspection and health assessment of the bearing capacity of pile foundations in these in-service bridges are essential to ensuring operational safety [2].

Traditional methods for evaluating the bearing capacity of pile foundations mainly rely on the static loading test (SLT) and high-strain dynamic pile testing (HSDPT). The SLT is regarded as the benchmark method for determining bearing capacity, and its results are intuitive and reliable. However, for in-service bridge inspection, the SLT faces inherent limitations, such as high cost, long testing duration, the need for large reaction systems, and stringent site requirements [3]. More importantly, the destructive nature of the SLT and its substantial spatial requirements make it difficult to apply on bridges under normal operation. Although HSDPT improves testing efficiency, the impact of heavy hammer blows may cause damage to the pile body, and its results still require validation through the SLT. Svinkin [4] pointed out that signal processing and CAPWAP analysis in HSDPT exhibit significant non-uniqueness under complex geological conditions. Likins and Rausche [5], based on extensive comparative studies, showed that the discrepancy between HSDPT and SLT results can reach ±20%. Furthermore, pile foundations of in-service bridges are usually integrated with pile caps and superstructures, making conventional HSDPT difficult to implement.

As a non-destructive testing approach, dynamic methods provide a physical basis for evaluating the bearing capacity of pile foundations [6]. Among these approaches, the dynamic stiffness index obtained from the transient response method (TRM) has a theoretical correlation with the static characteristics of the pile–soil system. Davis [7] provided a comprehensive review of applications of the TRM in North America, confirming its potential in foundation assessment. Lo et al. [8] investigated the response characteristics of in-service bridge substructures under transient loading through time–frequency-domain analysis. In recent years, Liu and Ma [9] conducted large-scale model tests to systematically study the dynamic stiffness characteristics of different types of foundations, demonstrating a significant positive correlation between vertical dynamic stiffness and the ultimate bearing capacity of foundations. Ma et al. [10] proposed a pile foundation bearing capacity estimation model based on the TRM, while Chu et al. [11] analyzed the dynamic interaction characteristics of pile cap–pile systems.

However, the application of the dynamic stiffness method still faces two major challenges. First, the measured dynamic stiffness is significantly influenced by human factors such as excitation energy, sensor coupling conditions, and the selection of signal frequency bands [12]. Second, the dynamic–static conversion coefficient η is affected by the complex coupling of geological conditions, pile dimensions, and nonlinear pile–soil interactions. In traditional approaches, the determination of η relies heavily on regional experience, leading to considerable uncertainty [13].

In recent years, machine learning (ML) has been increasingly applied in geotechnical engineering, providing a new pathway for addressing complex nonlinear mapping problems. Existing studies have shown that high-accuracy prediction models for bearing capacity can be developed using algorithms such as artificial neural network (ANN) and support vector machine (SVM) based on Standard Penetration Test (SPT) or Cone Penetration Test (CPT) data [14]. For example, Shahin [15] effectively modeled the load–settlement behavior of steel piles using recurrent neural networks, while Nejad and Jaksa [16] successfully captured the nonlinear response characteristics of single piles in complex soil layers using ANN. Recently, Cao et al. [17] proposed a rapid evaluation method for the vertical bearing capacity of pile group foundations using machine learning, and Elbaroty et al. [18]

successfully applied tree-based ensemble algorithms (e.g., XGBoost) to forecast bridge health degradation, further demonstrating the potential of such data-driven approaches in bridge foundation assessment. In addition, ensemble learning algorithms such as random forest (RF) and XGBoost [19,20] have been introduced into pile foundation design frameworks, significantly improving prediction accuracy. Furthermore, the weighted voting ensemble model proposed by Pham and Vu [21] further demonstrates the reliability of data-driven approaches in geotechnical engineering decision making.

However, most of the aforementioned studies rely on destructive or borehole sampling data obtained during the site investigation stage (such as CPT and SPT), making them difficult to apply directly to in-service bridges where original geotechnical data are often unavailable. Nevertheless, data-driven methods have achieved significant breakthroughs in the field of structural health monitoring. Bulut et al. [22] achieved real-time non-destructive monitoring by combining SVM with wavelet analysis, while Pan et al. [23] explored damage diagnosis of cable-stayed bridges based on time–frequency features. In terms of substructure assessment, Zhao et al. [24] quantitatively identified bridge hinge joint damage using SVM combined with differences in dynamic stiffness, and Ni et al. [25] proposed a rapid assessment method for pier damage based on longitudinal dynamic stiffness and the k-nearest neighbors (KNN) algorithm. Inspired by these studies, directly utilizing non-destructive dynamic features to construct an end-to-end classification model for bearing capacity compliance holds significant engineering application value.

To address the difficulty of evaluating the bearing capacity of in-service bridge pile foundations, this paper proposes a novel discrimination method based on dynamic stiffness indicators and ML. Unlike traditional methods that attempt to precisely solve the complex dynamic–static comparison coefficient η, this study directly utilizes the dynamic stiffness obtained through non-destructive means as input features, integrating pile geometry and load information, to construct a statistical mapping from input features $\boldsymbol{X}$ to engineering decision targets $\boldsymbol{Y}$ (whether the bearing capacity meets requirements). This approach aims to circumvent the uncertainty of η and to explore the deeper statistical correlation between K_d and bearing capacity, providing a data-driven solution for the rapid screening of in-service pile foundations.

2. Materials and Methods

2.1. Correlation Between Dynamic Stiffness and Bearing Capacity

In pile foundation engineering, dynamic stiffness K_d is a key indicator for evaluating the vertical dynamic characteristics of foundations. Physically, it originates from the linear elastic response of the pile–soil system in the low-frequency range. According to mechanical impedance theory, the admittance response of the system in the low-frequency range (typically < 50 Hz) varies linearly with frequency, the slope defines the compliance, and dynamic stiffness is its reciprocal.

It should be noted that dynamic stiffness reflects the system's behavior in the small-strain range. Although the ultimate bearing capacity Q_u is positively correlated with dynamic stiffness, the relationship is not a simple linear proportionality. Theoretically, the bearing capacity of a pile foundation should be determined by jointly considering the static stiffness K_s and the allowable settlement $[S_a]$, i.e., $Q_u = K_s \cdot [S_a]$. This relationship represents a simplified linear approximation that is more appropriate for friction piles, where load is transferred progressively along the pile shaft. For end-bearing piles that primarily rely on tip resistance, the load–settlement response may exhibit stronger nonlinearity, and the validity of this linear assumption may be reduced. Accordingly, the bearing mode (friction pile vs. end-bearing pile) is explicitly included as a feature in the machine learning model, allowing the model to learn the distinct mechanical behavior of each pile type.

Structural defects (e.g., necking, honeycombing, and sediment accumulation at the pile tip) primarily affect the structural integrity of the pile. However, such defects simultaneously reduce the overall stiffness of the pile–soil system, which can indirectly compromise the geotechnical bearing capacity; the latter being the primary evaluation target of static load tests (SLTs). Therefore, abnormally low dynamic stiffness measured by the TRM is treated in this study as a warning indicator of potential bearing capacity deficiency, based on the physical rationale that reductions in system stiffness and bearing capacity tend to occur simultaneously. Consequently, abnormally low values of K_d measured by the TRM can serve as an effective indicator of potential bearing capacity deficiency.

2.2. Dynamic Stiffness Testing Method and Data Acquisition

2.2.1. Testing Principle and Instrumentation

In this study, the TRM was employed to obtain dynamic stiffness. Given that the pile foundations of in-service bridges are typically integrated with pile caps and superstructures, which impose significant constraints on the testing system, sufficient excitation energy is required to effectively mobilize the overall dynamic response of the pile–soil system. To this end, the testing equipment utilized an automatic drop-weight system (KingSci Instruments, Beijing, China) developed by Ma et al. [1]. This system is designed to provide sufficient low-frequency excitation energy for in-service bridges while ensuring that no damage is caused to the structure. It should be acknowledged that the measured dynamic stiffness K_d represents the combined response of the pile-cap-soil system rather than an isolated single pile, due to the mechanical constraints imposed by the pile cap, adjacent piles, and the superstructure. While these constraints affect the absolute magnitude of the measured values, the relative variation in K_d across piles remains a reliable health indicator within a structurally consistent bridge inventory, as the systematic constraint effects are largely uniform across the bridges in this project. This has been demonstrated in Liu and Ma [9] and Ma et al. [10].

The testing system mainly consists of three components:

(1) Excitation device: The automatic drop-weight system consists of a 200 kg hammer with a lifting height of 1.5–2.0 m, capable of generating a transient impact force of up to 200 kN (Figure 1). A typical time history and Fourier spectrum of measured force is illustrated in Figure 2. A rubber cushion pad is attached to the hammer to ensure that the impact energy is concentrated in the low-frequency range below 50 Hz, which aligns with the frequency band (10–30 Hz) selected for dynamic stiffness calculation.

(2) Sensor system: A vertical velocity sensor is installed on the top surface of the pile cap to capture the velocity response of the pile–soil system, while a force sensor integrated at the top of the hammer is used to synchronously record the impact force time history $F(t)$.

(3) Data acquisition and analysis system: An INV3068P (China Orient Institute of Noise & Vibration, Beijing, China) intelligent signal acquisition system is used as the recording hardware, and a 941B vibration sensor (Institute of Engineering Mechanics, China Earthquake Administration) is employed to capture the velocity response of the pile–soil system. The impact force time history $F(t)$ is synchronously recorded by a force sensor integrated at the top of the hammer. All signals are processed using DASP data acquisition and processing software (Version V11, China Orient Institute of Noise & Vibration, Beijing, China). To improve repeatability and reliability, 3–5 repeated impacts are applied at each test location on the pile cap. Among these, three sets of data with high consistency and minimal waveform deviation are selected for averaging to reduce random measurement errors. Detailed signal processing procedures (e.g., windowing) can be found in Ref. [1].

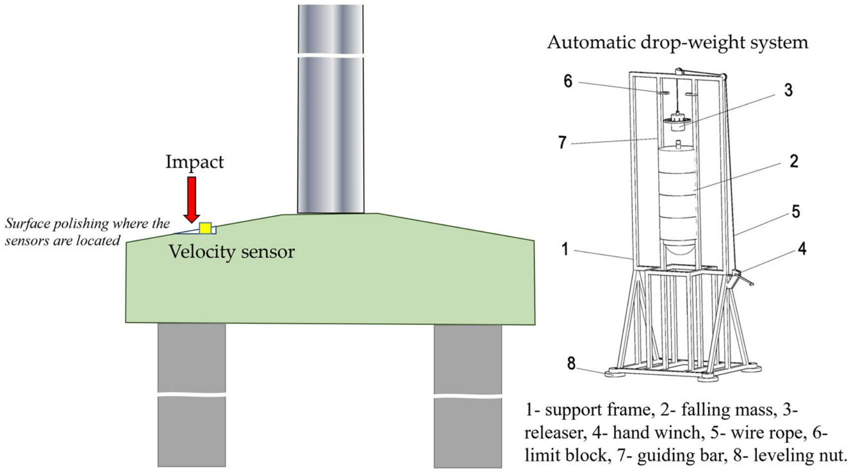

Figure 1. Schematic diagram of the TRM measurement setup.

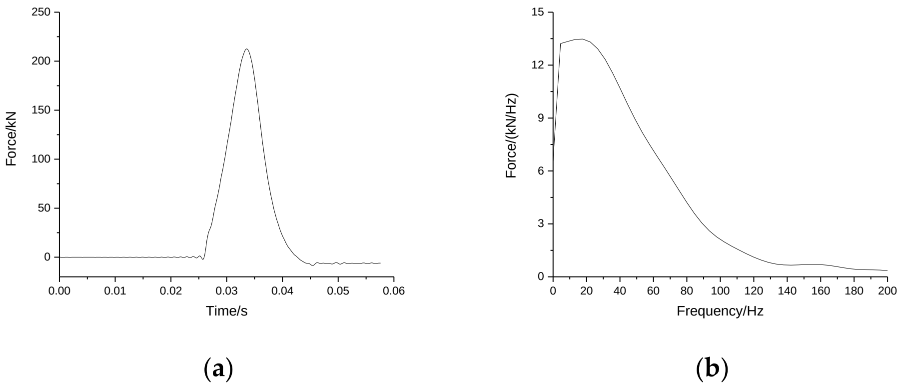

Figure 2. Typical (**a**) time history and (**b**) Fourier spectrum of measured force.

2.2.2. Signal Processing and Dynamic Stiffness Calculation

Under impact loading, the structural response initially exhibits pronounced transient vibrations, which then gradually decay over time. To mitigate the effects of signal truncation and spectral leakage on the calculation results, an exponential window function is applied to the velocity response signal to accelerate the decay of vibrations and ensure a smooth transition to zero. According to Ma et al. [1], the attenuation coefficient should be greater than 5 to effectively eliminate truncation effects, allowing the dynamic stiffness function $K_d(f)$ to stabilize in the low-frequency range.

Following windowing, the measured impact force time history $F(t)$ and velocity response time history $v(t)$ are transformed into the frequency domain by using the Fourier transform, yielding the corresponding spectra $F(f)$ and $V(f)$. Based on mechanical admittance theory, the ratio of these spectra, $V(f)/F(f)$, characterizes the dynamic admittance of the structure.

Accordingly, the dynamic stiffness $K_d(f)$ can be expressed as:

$$K_d(f) = \left| \frac{2\pi f}{V(f)/F(f)} \right| \tag{1}$$

In practical applications, the theoretical basis for dynamic stiffness calculation is illustrated in Figure 3, which shows a typical measured mobility spectrum obtained from a pile foundation using the TRM. In the low-frequency range (typically below 50 Hz), the measured mobility response exhibits an approximately linear variation with frequency, which is consistent with the theoretical behavior of a rigid body on an elastic foundation. The slope of this linear portion defines the compliance (flexibility) of the tested area under a normalized force input, and the inverse of this compliance gives the dynamic stiffness K_d of the structural element at the test point. Accordingly, the average dynamic stiffness within the frequency band of 10–30 Hz, where the signal energy is relatively strong and low-frequency excitation characteristics dominate, is selected as the vertical dynamic evaluation index for the pile foundation. This frequency band selection aligns with the excitation characteristics of the testing system (Section 2.2.1) and effectively avoids higher-frequency interference that may be unrelated to the overall stiffness of the pile–soil system.

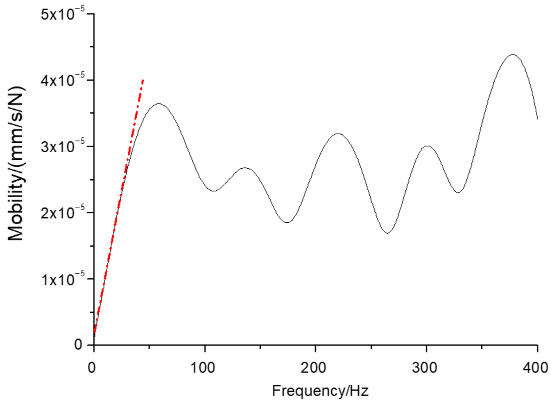

Figure 3. Measured mobility response spectrum obtained using the TRM. The red dash-dot line indicates the low-frequency asymptote for deriving dynamic stiffness.

2.3. Data Sources

The data used in this study were obtained from a special inspection project of pile foundations for a mega-bridge complex along an expressway in southern China. All piles in this dataset are cast-in-place bored concrete piles (drilled shafts), which are the dominant type for highway bridges in the studied region. Driven steel piles were not included. The expressway consists of 32 major bridges, and a total of 680 pile foundation samples were tested and collected during this project. As these bridges have been in service for several decades, a considerable number of pile foundations exhibit various defects and deterioration, necessitating inspection and potential reinforcement. Consequently, the dataset contains a substantial number of samples that do not meet the bearing capacity requirements, providing a representative basis for developing and validating the proposed classification model.

For each test sample, the following information was comprehensively recorded:

Input features include pile diameter D; designed pile length L (defined as the length from the bottom of the pile cap to the pile tip); design load P; measured dynamic stiffness K_d; pier section type (classified into three categories: A, B, and C; see Figure 4); and load-bearing mode (friction pile or rock-socketed pile). The calculated bearing capacity Q_{calc} according to design codes was also recorded, but it was not used as an input feature in the proposed machine learning model. Q_{calc} is derived from the empirical formula specified in Chinese code JTG D 63-2007 Code for Design of Ground Base and Foundation of Highway Bridges and Culverts [26]:

$$Q_{calc} = u \sum l_i q_{ik} + \alpha \cdot A \cdot q_{pk} \tag{2}$$

where u is the perimeter of the pile cross-section; l_i and q_{ik} are the thickness and ultimate side friction of the i-th soil layer, respectively; A is the pile tip area; q_{pk} is the ultimate tip resistance; and α is an empirical coefficient that implicitly accounts for the dynamic–static conversion effect under the framework of the static loading test (SLT) reference system. As highlighted by Ma et al. [1], the dynamic stiffness obtained via the TRM can effectively reflect the vertical condition of a bridge substructure, confirming its physical relevance as a health indicator for pile foundations. However, the reliance of conventional capacity estimation on the empirical coefficient α introduces considerable uncertainty, as α is highly dependent on local geological conditions, pile types, and construction quality, and typically requires calibration against SLT results.

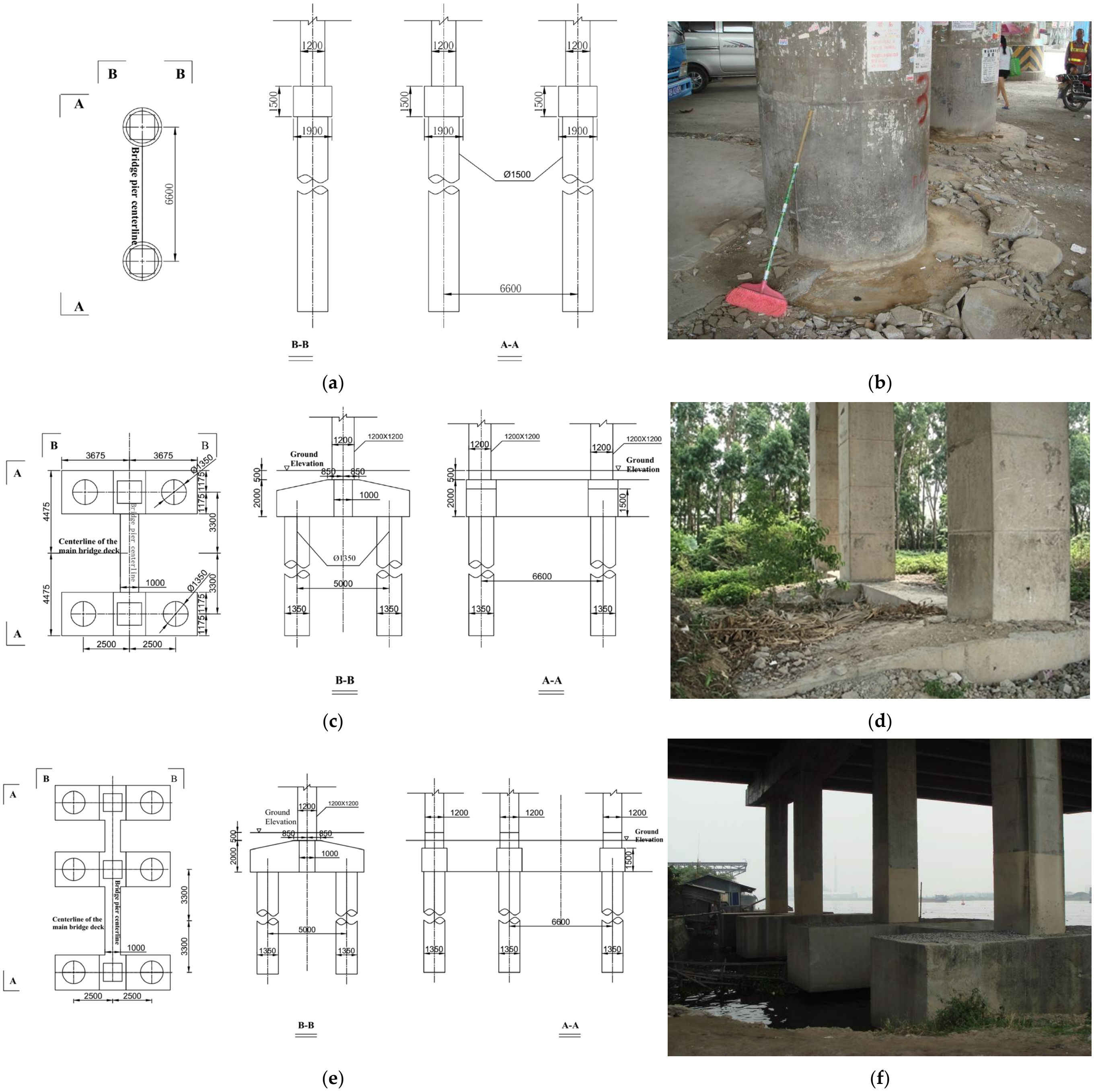

Figure 4. Three types of bridge pier sections (unit: mm). Type A: (**a**) schematic drawing and (**b**) photograph. Type B: (**c**) schematic drawing and (**d**) photograph. Type C: (**e**) schematic drawing and (**f**) photograph.

To circumvent this issue, the proposed machine learning model does not use Q_{calc} as an intermediate predictor. Instead, it directly maps the raw physical features $\boldsymbol{X} = [K_d,\ D,\ L,\ P,$ "pile type", "load-bearing mode"$]$ to the final binary decision $\boldsymbol{Y}$ (meeting/not meeting design requirements). By establishing a statistical mapping between the non-destructively measured dynamic stiffness and the engineering decision target, this approach bypasses the uncertainty introduced by the empirical dynamic–static conversion coefficient while leveraging the proven effectiveness of dynamic stiffness as a condition indicator [1].

Decision label $\boldsymbol{Y}$ was determined by an expert panel based on a comprehensive evaluation of Q_{calc}, low-strain integrity testing results, and 5% core drilling verification. The label is binary, classified as "meeting design requirements (1)" or "not meeting requirements (0)".

The binary label necessarily involves expert judgment, because full-scale destructive static load tests are infeasible for in-service bridges. This multi-source expert synthesis is currently the accepted engineering standard for assessing in-service pile foundations and is therefore the best available approximation of the true bearing capacity status. The potential impact of label noise (occasional mislabeling) on model performance is 2-fold: it lowers the achievable upper bound of classification accuracy and may cause the model to learn spurious patterns. Nevertheless, the fact that all three models, including the simple linear logistic regression, achieved accuracy exceeding 95% suggests that the underlying physical signal (particularly dynamic stiffness K_{d}) is strong and that label noise is not dominant. Moreover, ensemble methods such as random forest and gradient boosting decision tree (GBDT) are known to be relatively robust to moderate levels of label noise. If static load test results become available for a subset of piles in future work, they would provide a more definitive validation benchmark.

Representative core samples illustrating the correlation between dynamic stiffness and bearing capacity are presented in Figure 5. Pile No. D11-E2, with an extremely low dynamic stiffness of 2.8×10^9 N/m, exhibited severe segregation and honeycombing defects and was consequently assessed as "not meeting requirements". In contrast, Pile No. N2-C1, with a dynamic stiffness of 4.146×10^9 N/m, showed only minor local defects and was assessed as "meeting requirements". These visual observations provide direct physical evidence supporting the premise that low dynamic stiffness reflects underlying stiffness deficiencies that compromise bearing capacity.

(**a**)

(**b**)

Figure 5. Representative core samples from piles with different bearing capacity assessments: (**a**) Pile No. D11-E2, which failed to meet requirements, showing severe segregation and honeycombing. (**b**) Pile No. N2-C1, which met requirements despite minor defects (local honeycombing and tip sediment).

The core samples serve as visual validation: Severe segregation/honeycombing (Pile D11-E2) directly reduces the cross-sectional stiffness and strength, leading to lower dynamic stiffness and inadequate bearing capacity. In contrast, minor local defects (Pile N2-C1) do not significantly affect overall capacity. Thus, the core photos support the correlation between dynamic stiffness and geotechnical bearing capacity, not just concrete integrity.

Table 1 presents the statistical description of the dataset. It can be observed that the numbers of samples classified as "meeting requirements" (N = 332) and "not meeting requirements" (N = 348) are approximately balanced. The ratio between the two classes is close to 1:1, which is beneficial for the model to learn an unbiased decision boundary and helps avoid potential bias caused by class imbalance.

Table 1. Statistical description of key features in the dataset.

Feature	Mean	Standard Deviation	Minimum	Maximum
Pile diameter D (m)	1.41	0.11	1.2	1.5
Designed pile length L (m)	24.87	8.51	2.23	45
Applied load P (kN)	5921.5	2197.8	3918	10,330
Dynamic stiffness K_d ($\times 10^9$ N/m)	6.48	3.01	2.5	17.86
Calculated bearing capacity Q_{calc} (kN)	6962.3	2925.4	2148	15,333
Number of samples (Pass/Fail)	680	332 (Pass)/348 (Fail)	-	-

2.4. Data Cleaning and Preprocessing

To meet the input requirements of machine learning models, the raw data were preprocessed as follows:

(1) Data integrity was first checked, and samples with obvious logical errors or missing key fields were removed. For example, four piles had missing records for "pile diameter", and two piles had severe noise interference in the dynamic stiffness signals, making them unusable; these samples were therefore excluded. A total of 674 valid samples were ultimately retained.

(2) Categorical features in textual form were converted into numerical variables. The "bearing mode" was encoded using one-hot encoding, generating two binary features: "friction pile" and "rock-socketed pile" (with end-bearing pile as the reference category). For the "pier section type", which is a multi-category feature (e.g., A/B/C), preliminary analysis showed strong collinearity with "bearing mode" and geometric parameters. Moreover, direct one-hot encoding would significantly increase the feature dimensionality (expanding to 3 dummy variables for 674 samples). To avoid the curse of dimensionality and overfitting, this feature was excluded as an independent input in this study and is reserved for future research using more advanced embedding techniques.

(3) A stratified sampling approach was adopted to divide the dataset into a training set (80%, 539 samples) and an independent test set (20%, 135 samples), maintaining the original proportion of "meeting" and "not meeting" classes. The test set was held out entirely during the entire training and hyperparameter optimization process and was used exclusively for final model evaluation to ensure an unbiased assessment of model performance.

(4) To eliminate the influence of different scales and magnitudes among features, particularly for models such as logistic regression and distance-based algorithms, all numerical features (dynamic stiffness K_d, pile length L, pile diameter D, and applied load P) were standardized using Z-score normalization, resulting in a mean of 0 and a standard deviation of 1. The standardization parameters (mean μ and standard

deviation σ) were computed exclusively on the training set and then applied to both the training and test sets to prevent data leakage.

The complete data preprocessing workflow is summarized in Figure 6.

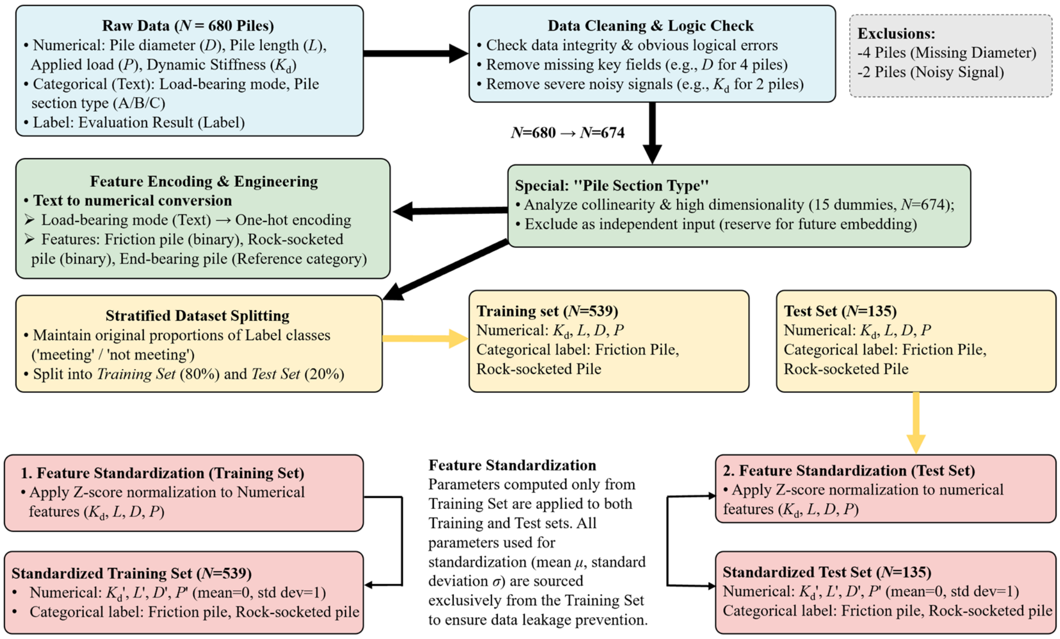

Figure 6. Flowchart of the data preprocessing procedure.

3. Machine Learning Model Construction and Optimization

3.1. Problem Reformulation: From Multi-Class Classification to Bearing Capacity Qualification Assessment

In the initial stage of the present research study, an attempt was made to construct a four-class prediction model based on existing specifications, aiming to categorize pile foundations into Category I (high-quality piles), Category II (good piles), Category III (qualified but with defects, suitable for service under conditions), and Category IV (unqualified piles). However, preliminary experimental results showed that the overall accuracy of this model on the validation set was only about 40%, and with notable predictive confusion observed between Category II and Category III samples.

An in-depth analysis suggested that the classification failure stemmed primarily from two aspects. Firstly, the distinction between Categories II and III involves a certain degree of subjectivity and a gray area in engineering definition, as its determination often relies on qualitative indicators such as waveform integrity, resulting in severe overlap among samples in the feature space. Secondly, given the total sample size of 680 pile foundations, discretizing the continuously varying bearing capacity condition into four levels introduces inter-class complexity and potential imbalance, making it difficult for algorithms to learn a robust decision boundary.

To address these limitations, this study refocused on the core decision-making problem in engineering practice: "Does the bearing capacity of the pile foundation meet the design requirements?" Accordingly, the prediction target was reformulated as a binary classification problem: the original Categories I and II were merged into positive samples (meeting requirements), and the original Categories III and IV were merged into negative samples (not meeting requirements). This reformulation not only aligns the model's optimization objective directly with the engineer's decision-making logic, namely, determining whether the bearing capacity meets the design requirements, but also ensures that the physical essence of the two classes (i.e., whether the overall stiffness of the pile–soil system is sufficient) is more clearly distinct, thereby laying a solid foundation for subsequent breakthroughs in model performance.

A three-class scheme (meets/doubtful/fails) was considered but not adopted in the current study because the engineering decision logic in practice requires a clear binary "pass/fail" threshold. Moreover, the "doubtful" samples often correspond to the fuzzy boundary between the original Categories II and III, which contributed to the poor performance of the four-class model. Future work may introduce a reject option or confidence-based classification to handle borderline cases.

3.2. Machine Learning Algorithm Framework

To comprehensively evaluate the performance of the binary classification task, this study selected three representative machine learning algorithms for comparative experiments.

3.2.1. Logistic Regression (LR)

As a representative of generalized linear models, logistic regression was used as the baseline model for this study. It assumes a linear relationship between the features $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$ and the log-odds, and maps the linear combination to a probability value through the sigmoid function. Its mathematical expression is

$$\boldsymbol{P}(y = 1 \mid \boldsymbol{X}) = \frac{1}{1 + \mathbf{e}^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_\mathrm{n} x_\mathrm{n})}} \tag{3}$$

where $P(y = 1 \mid \boldsymbol{X})$ represents the probability that a sample belongs to the positive class (bearing capacity meets the requirements) given features $\boldsymbol{X}$, and β_i denotes the weight coefficients to be learned. The model estimates the parameters β_i by maximizing the likelihood function. The weight coefficients of logistic regression can directly reflect the direction and magnitude of the influence of features on the outcome, providing strong interpretability. This study uses it as the baseline model to test whether the judgment of bearing capacity qualification is linearly separable in the current feature space.

3.2.2. Random Forest (RF)

Random forest is an ensemble learning algorithm based on the Bagging strategy. It is constructed through the following steps: First, in bootstrap sampling, randomly draw k sample subsets with replacement from the original training set to construct T decision trees. Second, in random feature selection, at each node split of each decision tree, randomly select m_{try} features from the total M features (typically $m_{try} \approx \sqrt{M}$), and choose the optimal split feature from them. Third, in ensemble voting, for classification tasks, the final prediction result is determined by the vote of all decision trees:

$$\hat{\boldsymbol{y}} = \text{majority vote}\{\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \ldots, \hat{\boldsymbol{y}}_{\boldsymbol{T}}\} \tag{4}$$

By introducing sample perturbation and feature perturbation, random forest effectively reduces the risk of overfitting associated with a single decision tree and exhibits strong robustness to outliers and noise. Additionally, during the training process, the model can calculate the average reduction in impurity (such as Gini impurity) across all decision trees for each feature, serving as a feature importance score. This study uses it as a strong nonlinear baseline model to capture the potentially complex interaction effects between dynamic stiffness and bearing capacity.

3.2.3. Gradient Boosting Decision Tree (GBDT)

GBDT is an ensemble learning algorithm based on the Boosting strategy. Unlike random forest, which constructs trees in parallel, GBDT uses a sequential, iterative training approach. The core idea is to train a new decision tree in each iteration to fit the residuals (i.e., the negative gradient direction) of the predictions from the previous tree, thereby gradually reducing the overall bias. Its mathematical expression can be summarized as

$$\boldsymbol{F_m}(\mathbf{x}) = \boldsymbol{F_{m-1}}(\mathbf{x}) + \boldsymbol{\nu} \cdot \boldsymbol{h_m}(\mathbf{x}) \quad (5)$$

where $F_m(\mathrm{x})$ is the ensemble model after m iterations, $h_m(\mathrm{x})$ is the new decision tree trained in the current iteration (used to fit the residuals), and v is the learning rate (shrinkage parameter), which controls the contribution of each tree to the final result to prevent overfitting. GBDT typically achieves a higher accuracy ceiling than random forest but is more sensitive to hyperparameters (such as learning rate, tree depth, and number of iterations), requiring careful tuning. This study employs it to pursue optimal predictive performance, aiming to provide the most reliable screening tool for practical engineering.

3.3. Model Training, Hyperparameter Optimization, and Evaluation Metrics

This study used Python's Scikit-learn library (version 1.2.0) for model construction and evaluation. Before model training, all numerical features (dynamic stiffness K_d, pile length L, pile diameter D, and applied load P) in the training set (539 samples) were standardized using Z-score normalization. The mean and standard deviation from the training set were then applied to transform the test set (135 samples) to prevent data leakage.

To achieve optimal model performance and prevent overfitting, 5-fold cross-validation combined with grid search was employed within the training set for hyperparameter tuning. In this process, the training set is split into five folds; each iteration uses four folds for training and the remaining fold as an internal validation set to evaluate performance. The evaluation metric for this internal validation was the F1 score, to balance precision and recall, with particular attention to the model's ability to identify samples "not meeting the requirements". This internal validation set is distinct from the independent test set described in Section 2.4; the latter was held out entirely and used only for final model evaluation. The main hyperparameter search ranges for each model were set as follows:

- Logistic Regression: Regularization strength C was searched within {0.01, 0.1, 1, 10, 100}, using L2 regularization.
- Random Forest: Number of trees n_estimators was searched within [50, 100, 200], maximum depth max_depth within {5, 10, 15, None} and minimum samples per leaf min_samples_leaf within {1, 2, 4}.
- GBDT: Learning rate learning_rate was searched within [0.01, 0.05, 0.1], number of iterations n_estimators within {100, 200, 300}, and maximum depth max_depth within {3, 5, 7}.

Finally, the parameter combination achieving the highest average F1 score during cross-validation was selected as the final parameters for that model. The model was then retrained on the entire training set and finally evaluated on the independent test set.

To comprehensively measure model performance, this study adopted the following metrics:

- Accuracy: The proportion of correctly predicted samples out of the total number of samples.
- Precision: Among the samples predicted as "meeting requirements", the proportion that actually "meet requirements".
- Recall: Among the samples that actually "meet requirements", the proportion correctly predicted (for the positive class).
- F1 Score: The harmonic mean of precision and recall.
- AUC Value: The area under the Receiver Operating Characteristic curve, reflecting the model's ability to distinguish between positive and negative classes.

4. Result Analysis and Discussion

4.1. Model Performance Comparison and Analysis

After 5-fold cross-validation and grid search optimization, the performance of the three models, logistic regression (LR), random forest (RF), and gradient boosting decision tree (GBDT), on the independent test set is shown in Table 2. To intuitively present the differences among algorithms across multiple metrics, Figure 7 displays radar charts of each model on four metrics, accuracy, precision, recall, and F1 score, while Figure 8 presents the confusion matrix visualization results for each model.

Table 2. Performance comparison of different models on the test set.

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	95.60%	0.96	0.95	0.95	0.99
Random Forest	95.60%	0.97	0.94	0.95	0.98
Gradient Boosting Decision Tree	96.30%	0.97	0.96	0.96	0.99

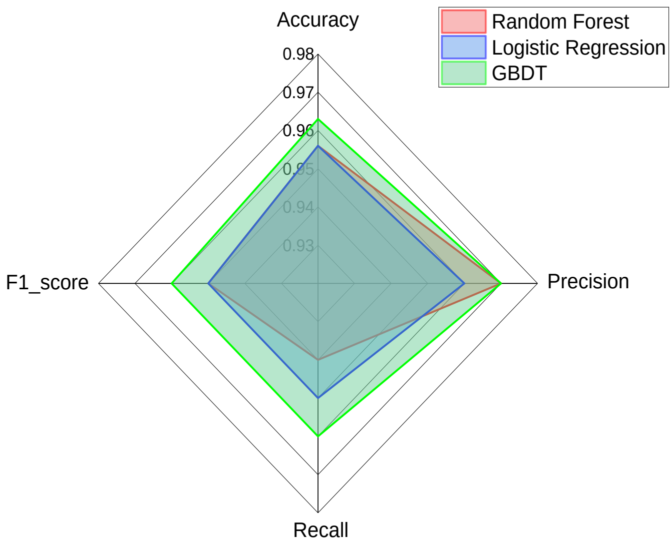

Figure 7. Radar chart of performance metrics for three models.

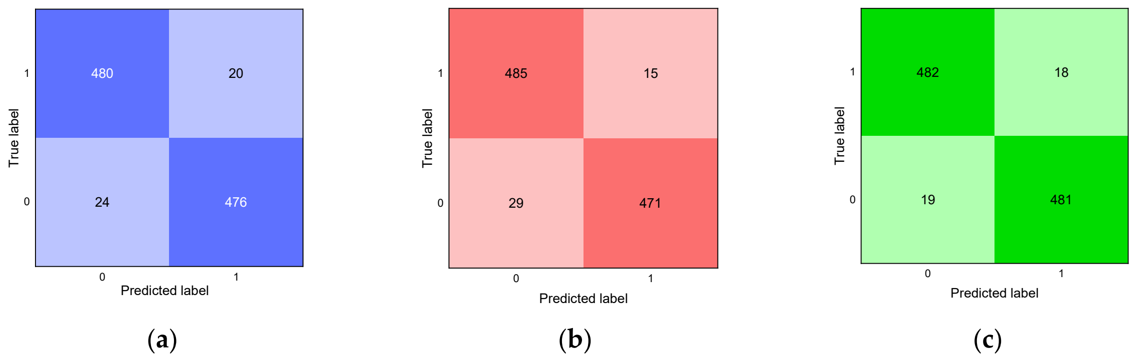

Figure 8. Confusion matrix visualization: (**a**) logistic regression, (**b**) random forest and (**c**) GBDT.

The experimental results demonstrate that all three models achieved excellent performance on the binary classification task. After reformulating as a "bearing capacity qualification" problem, each model attained accuracy rates exceeding 95%, with AUC values approaching 1, indicating that clear decision boundaries can be learned from the available features. This finding validates the effectiveness of aligning the learning objective with the engineering decision objective.

Notably, logistic regression, a linear model, achieved nearly equivalent performance to the ensemble models (accuracy of 95.6% and F1 score of 0.95). This result carries significant engineering implications, as it suggests that in the current feature space, the decision boundary between "meets requirements" and "does not meet requirements" samples exhibits good linear separability. In other words, features such as dynamic stiffness and pile length, through appropriate linear weighted combinations, can already effectively distinguish between the two categories of pile foundations. From a data-driven perspective, this strongly corroborates the physical hypothesis that low dynamic stiffness directly correlates with insufficient bearing capacity, because if the correlation were nonlinear and complex, linear models would struggle to achieve such high performance.

Further analysis reveals that the GBDT model is particularly well suited for safety-critical applications. While overall performance metrics were comparable across models, GBDT achieved the highest recall (0.96), outperforming logistic regression (0.95) and random forest (0.94). In the context of engineering safety, however, greater attention should be paid to the model's ability to correctly identify samples that do not meet requirements, i.e., the true negative rate. Examination of the confusion matrices (Figure 8) shows that GBDT misclassified only one "does not meet" sample as "meets", whereas logistic regression and random forest had two and three such missed detections, respectively. In pile foundation safety screening, the cost of missed detections (classifying non-compliant piles as compliant) far exceeds that of false alarms (classifying compliant piles as non-compliant), because missed detections may lead to overlooked safety hazards. Consequently, from an engineering application perspective, GBDT offers the greatest practical value due to its lowest missed detection rate.

To further verify the necessity of the problem reformulation, the performance of the binary classification models was compared with that of the initial four-class model. The four-class model achieved only 37.5% overall accuracy on the validation set, with confusion rates exceeding 45% between Class II and Class III samples. This stark contrast underscores that when the classification scheme is aligned with engineering decision objectives, the predictive potential of the model is fully unleashed.

4.2. Feature Importance Analysis

The feature importance rankings derived from the GBDT model are presented in Figure 9. Since the importance rankings from both the random forest and GBDT models were highly consistent, only the GBDT results are shown here as a representative example.

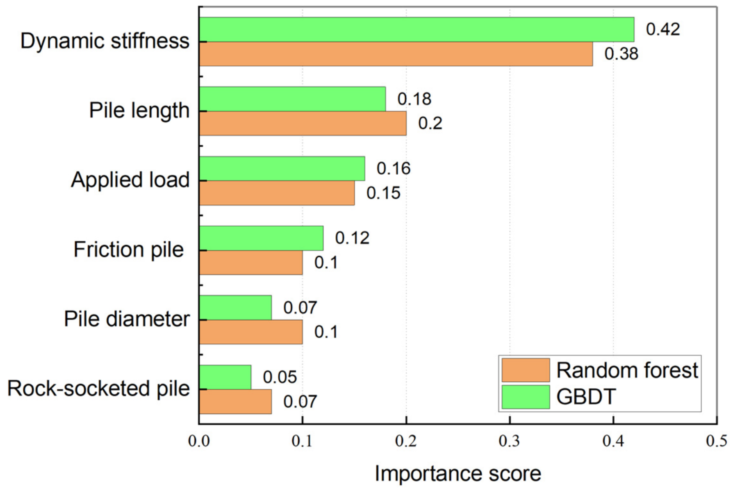

Figure 9. Feature importance ranking based on GBDT.

Analysis of Figure 9 reveals the following:

Across all models, the importance score of "dynamic stiffness" is the highest among all features, contributing approximately 42% to the classification outcome (feature importance values are normalized to sum to 1). The remaining 58% is distributed among pile length, applied load, pile type, and other low-importance features. This result powerfully validates the core theoretical hypothesis of this paper in a data-driven manner: dynamic stiffness is a key indicator characterizing the overall stiffness of the pile–soil system, and its decrease directly reflects potential stiffness deficiencies in the system, which are highly likely to simultaneously threaten bearing capacity safety. From an engineering application perspective, this finding implies that in future rapid screening practices, if resources are limited, priority should be given to ensuring accurate acquisition of dynamic stiffness data; any method attempting to evaluate pile foundation bearing capacity should treat dynamic stiffness as the primary consideration.

Pile length and load reflect a "supply–demand" relationship: The importance of "design pile length" and "applied load" follows closely, accounting for 18% and 16%, respectively. Pile length directly affects the mobilization of pile shaft friction resistance and is a key geometric parameter determining the "supply" capacity of friction pile bearing capacity; applied load represents the "demand" of the superstructure on the pile foundation. Whether the bearing capacity meets requirements is essentially a comparison of "supply" and "demand". The model automatically captured this engineering logic: longer piles generally imply higher bearing capacity supply, while higher loads imply greater demand, and the trade-off between the two is crucial to the final determination.

The binary feature "is friction pile" also demonstrated stable contribution (12%), indicating that distinguishing between friction piles and rock-socketed piles, two different bearing mechanisms, is helpful for the model to make correct judgments. Friction piles primarily rely on shaft friction resistance, while rock-socketed piles also have end-bearing contributions; the two have different failure modes, and the model learned this difference.

Pile diameter exhibited relative low importance (7%). The limited variation range of pile diameter in this dataset (mainly 1.2 m and 1.5 m, with a standard deviation of only 0.11) may be the primary reason for its insignificant contribution. In datasets with broader pile diameter variation, the importance of this feature may increase.

Notably, the feature importance ranking is highly consistent with geotechnical mechanics theory. This indirectly indicates that although machine learning models are "black boxes", the patterns they learn are consistent with physical laws, enhancing the credibility of the model.

4.3. Error Case Analysis

Despite the overall model accuracy exceeding 95%, a small number of samples were still misclassified. In-depth analysis of these error cases helps to understand the limitations of the model and provides guidance for subsequent improvements. Figure 10 shows the distribution of all misclassified samples in the dynamic stiffness–pile length feature space, while Figure 11 presents the feature radar charts of two representative misclassified samples.

The first misclassification case (false positive) involves Pile No. D11-E2, which was predicted as "meets requirements" but did not actually meet them. This pile exhibited extremely low dynamic stiffness (2.8×10^9 N/m, more than one standard deviation below the mean), and pile integrity testing also indicated severe defects. As shown in Figure 10, however, this pile had a relatively long design pile length (28.21 m, above the mean). It appears that the model may have over-relied on the heuristic that "long pile imply high bearing capacity", while underestimating the severe stiffness deficiency revealed by the exceptionally low dynamic stiffness. This case reveals a limitation of the current model: when multiple features point to contradictory conclusions, the model's trade-off mechanism remains suboptimal. From an engineering perspective, this misclassification represents a dangerous outcome (missed detection), suggesting that in practical application, a manual review mechanism should be established for long piles with abnormally low dynamic stiffness.

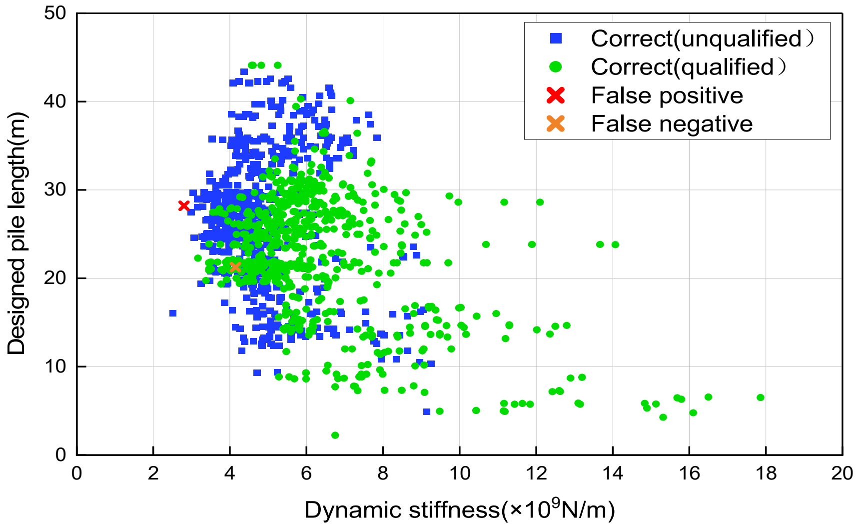

Figure 10. Scatter plot of feature distribution for misclassified samples.

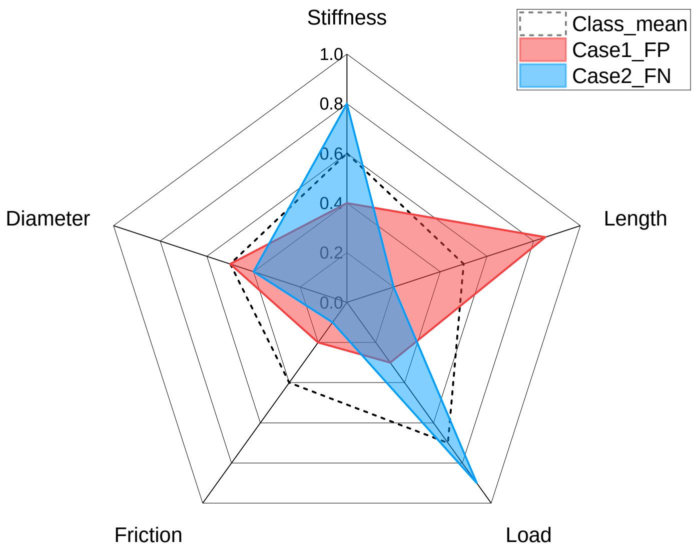

Figure 11. Feature radar charts of two typical misclassified samples.

The second misclassification case (false negative) concerns Pile No. N2-C1, which was predicted as "does not meet requirements" but actually met them. This rock-socketed pile had a dynamic stiffness value of 4.146×10^9 N/m, placing it in a "gray zone" close to but slightly below the mean of the "meets requirements" class. However, its calculated bearing capacity reached 7210 kN, far exceeding the applied load (3918 kN), leading to a final assessment of "meets requirements". The model likely made a conservative judgment based on the relatively low dynamic stiffness. As shown in Figure 11, this pile's applied load was substantially lower than that of similar samples, a key reason it actually met requirements, yet the model may not have given sufficient weight to this feature. This type of error is biased toward safety and falls within an acceptable range for engineering application, though excessive occurrences would reduce screening efficiency.

Notably, both of these misclassified samples were categorized as Class II or Class III in the preliminary four-class model, and the model exhibited low confidence in its prediction for them at that time. This observation indicates that even with high accuracy in the binary classification task, samples near the decision boundary (e.g., those with dynamic stiffness around a critical threshold) remain susceptible to misclassification. Such samples represent the "gray zone" that warrants focused attention in future work.

The above error cases highlight the importance of introducing additional features and employing interpretability analysis. For instance, introducing quantitative geological indicators (such as soil shear wave velocity), more refined descriptions of pile integrity (such as defect location and severity), or composite features like the "stiffness-to-load ratio" could further improve the model's ability to discriminate boundary samples. Moreover, subsequent studies could adopt interpretability tools such as SHAP to quantify the contribution of each feature to individual predictions, thereby providing engineers with intuitive explanations of the model's decision-making process.

4.4. Limitations and Future Work

Although the proposed GBDT-based binary classification model achieves high accuracy and a low false-negative rate on the test set, several limitations should be acknowledged, and corresponding future research directions are discussed below:

The current model is purely data-driven and does not embed governing physical equations (e.g., the wave equation). Future work will explore physics-informed neural networks (PINNs) to enforce physical consistency. Additionally, all data were collected from a single expressway corridor in southern China (weathered granite and alluvial soils); the model may not directly transfer to other geological settings. A multi-region database is being constructed, and local retraining or domain adaptation for new sites is recommended.

The binary "meets/does not meet" classification discards prediction confidence. A three-class scheme ("meets"/"doubtful"/"fails") or a probabilistic output with a rejection option will be developed to flag ambiguous piles.

As discussed in Section 2.3, ground-truth labels rely on expert judgment, which introduces subjectivity. Although inter-rater agreement was substantial, borderline cases remain ambiguous. Future work will adopt soft labels or probabilistic labeling based on multiple experts, and quantify aleatoric and epistemic uncertainty using Bayesian neural networks or ensemble methods.

In summary, while the current model is suitable for safety screening in the studied region, the above extensions will improve its physical consistency, robustness, predictive capability, and transferability.

5. Conclusions

Based on field measurement data from 680 bridge pile foundations, this study integrated dynamic stiffness indicators, geometric parameters, and design loads to construct machine learning models for rapid discrimination of pile foundation bearing capacity compliance. By comparing the performance of different algorithms and analyzing feature importance, the following main conclusions are drawn:

Firstly, reformulating the pile foundation bearing capacity evaluation problem from the traditional four-class classification to a binary classification task proved to be a critical step that aligns the modeling objective with engineering decision making. The binary classification models significantly outperformed the four-class model, demonstrating the feasibility of an end-to-end mapping paradigm that circumvents the uncertainty associated with conventional dynamic–static conversion coefficients.

Secondly, feature importance analysis shows that dynamic stiffness is the most critical feature for predicting bearing capacity compliance, with its contribution far exceeding factors such as pile length and pile type. This finding powerfully confirms the intrinsic physical correlation between dynamic stiffness and bearing capacity in a data-driven manner; low dynamic stiffness directly reflects potential stiffness deficiencies in the pile–soil system, thereby threatening bearing capacity safety. The proposed method provides a snapshot assessment of current bearing capacity. To predict dangerous trends under seasonal changes (e.g., freeze–thaw cycles in highland regions) or extreme events (e.g., floods), the long-term monitoring of dynamic stiffness combined with time-series models (e.g., long short-term memory, LSTM) would be required. This study lays the foundation for such future developments by identifying dynamic stiffness as the key indicator.

Thirdly, from an engineering perspective, all three binary classification models maintained accuracy above 95% on the test set, with the GBDT model performing the best (accuracy of 96.3% and F1 score of 0.96) and achieving the lowest missed detection rate, meeting the precision requirements for the rapid, non-destructive screening of in-service bridge pile foundations and demonstrating the potential for large-scale deployment in highway and railway maintenance monitoring programs.

Author Contributions: The authors confirm contributions to this paper as follows: conceptualization, M.M.; formal analysis, Y.Z., J.G., W.H. and Y.C.; investigation, Y.Z., J.G. and W.H.; methodology, Y.C. and M.M.; supervision, M.M.; writing—original draft Y.Z. and J.G.; writing—review and editing, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research study received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ma, M.; Liu, J.; Zhang, Z. Evaluating Vertical Conditions of Bridge Substructures of Heavy-Haul Railway Lines Based on Dynamic Stiffness and Pier Vibration Response. *Eng. Struct.* **2021**, *235*, 112037. [CrossRef]
2. Oh, H.-J.; Park, J.-H.; Park, H.-C. Non-Destructive Evaluation of Material Stiffness beneath Pile Foundations Tip Using Harmonic Wavelet Transform. *Buildings* **2024**, *14*, 511. [CrossRef]
3. Xu, X.; Zhu, P.; Song, Y.; Wang, C. Comparison of Load Transfer Law of Pipe Pile between O-Cell Test and Traditional Static Load Test. *Water* **2024**, *16*, 1245. [CrossRef]
4. Svinkin, M.R. High-Strain Dynamic Pile Testing-Problems and Pitfalls. *J. Perform. Constr. Facil.* **2010**, *24*, 99–108. [CrossRef]
5. Likins, G.E.; Rausche, F. Correlation of CAPWAP with Static Load Tests. In Proceedings of the Seventh International Conference on the Application of Stresswave Theory to Piles, Petaling Jaya, Malaysia, 3–5 August 2004; pp. 153–165.
6. Ren, J.; Sun, X. Prediction of Ultimate Bearing Capacity of Pile Foundation Based on Two Optimization Algorithm Models. *Buildings* **2023**, *13*, 1242. [CrossRef]
7. Davis, A.G. The Nondestructive Impulse Response Test in North America: 1985–2001. *NDT E Int.* **2003**, *36*, 185–193. [CrossRef]
8. Lo, K.F.; Ni, S.H.; Huang, Y.H. Non-Destructive Test for Pile Beneath Bridge in the Time, Frequency, and Time-Frequency Domains Using Transient Loading. *Nonlinear Dyn.* **2010**, *62*, 349–360. [CrossRef]
9. Liu, J.; Ma, M. A Model Test of the Dynamic Stiffnesses and Bearing Capacities of Different Types of Bridge Foundations. *Appl. Sci.* **2022**, *12*, 4951. [CrossRef]
10. Ma, M.; Liu, J.; Ke, Z.; Gao, Y. Bearing Capacity Estimation of Bridge Piles Using the Impulse Transient Response Method. *Shock Vib.* **2016**, *2016*, 4187026. [CrossRef]
11. Chu, J.H.; Ma, M.; Liu, J.B. Analysis of Dynamic Stiffness of Bridge Cap-Pile System. *Shock Vib.* **2018**, *2018*, 764572. [CrossRef]
12. Nishimura, A. Examination of Bridge Substructure for Integrity. In *Proceedings of the International Symposium on Traffic Induced Vibration Controls*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 131–142.
13. Zhan, J.W.; Xia, H.; Yao, J.B. Damage Evaluation of Bridge Foundations Considering Subsoil Properties. In *Environmental Vibrations: Prediction, Monitoring, Mitigation and Evaluation*; CRC Press: Boca Raton, FL, USA, 2005; pp. 271–277.

14. Sun, G.; Hasanipanah, M.; Amnieh, H.B.; Foong, L.K. Feasibility of Indirect Measurement of Bearing Capacity of Driven Piles Based on a Computational Intelligence Technique. *Measurement* **2020**, *156*, 107577. [CrossRef]
15. Shahin, M.A. Load–Settlement Modeling of Axially Loaded Drilled Shafts Using CPT-Based Recurrent Neural Networks. *Int. J. Geomech.* **2014**, *15*, 04014035. [CrossRef]
16. Nejad, F.P.; Jaksa, M.B. Load-Settlement Behavior Modeling of Single Piles Using Artificial Neural Networks and CPT Data. *Comput. Geotech.* **2017**, *89*, 9–21. [CrossRef]
17. Cao, Y.; Ni, J.; Chen, J.; Geng, Y. Rapid Evaluation Method to Vertical Bearing Capacity of Pile Group Foundation Based on Machine Learning. *Sensors* **2025**, *25*, 1214. [CrossRef] [PubMed]
18. Elbaroty, M.G.; Zaki, M.A.; Mourad, S.A. BMS Forecasting of Bridge Health Condition Degradation Using AI Machine Learning. *Civ. Environ. Eng.* **2025**, *21*, 246–253. [CrossRef]
19. Amjad, M.; Ahmad, I.; Ahmad, M.; Wróblewski, P.; Kamínski, P.; Amjad, U. Prediction of Pile Bearing Capacity Using XGBoost Algorithm: Modeling and Performance Evaluation. *Appl. Sci.* **2022**, *12*, 2126. [CrossRef]
20. Zhang, J.; Hu, J.; Li, X.; Li, J. Bayesian network based machine learning for design of pile foundations. *Autom. Constr.* **2020**, *118*, 103295. [CrossRef]
21. Pham, T.A.; Vu, H.L.T. Application of Ensemble Learning Using Weight Voting Protocol in Pile Bearing Capacity Prediction. *Math. Probl. Eng.* **2021**, *2021*, 6636734. [CrossRef]
22. Bulut, A.; Singh, A.K.; Shin, P.; Fountain, T.; Jasso, H.; Yan, L.J.; Elgamal, A. Real-Time Nondestructive Structural Health Monitoring Using Support Vector Machines and Wavelets. *Adv. Sens. Technol. Nondestruct. Eval. Struct. Health Monit.* **2005**, *5770*, 180–189.
23. Pan, H.; Azimi, M.; Yan, F.; Lin, Z. Time-Frequency-Based Data-Driven Structural Diagnosis and Damage Detection for Cable-Stayed Bridges. *J. Bridge Eng.* **2018**, *23*, 06018001. [CrossRef]
24. Zhao, J.; Zhan, J.; Wang, C.; Zhang, F.; Wang, Z.; Sun, Q.; Xu, X. A Dynamic Stiffness-Based Two-Step Method for Damage Identification of Joints in Hinged Slab Bridges Using Support Vector Machine. *Structures* **2023**, *58*, 105657. [CrossRef]
25. Ni, Z.; Zhan, J.; Xu, X.; Wang, C.; Wang, Z.; Wang, Y. A Rapid Damage Identification Method for Highway Bridge Double-Column Piers Based on Longitudinal Dynamic Stiffness. *Eng. Fail. Anal.* **2025**, *182*, 110192. [CrossRef]
26. *JTG D63-2007*; Code for Design of Ground Base and Foundation of Highway Bridges and Culverts. China Communications Press: Beijing, China, 2007.