https://doi.org/10.15302/J-LAF-0-020023 Received Date | 2023-09-03 Accepted Date | 2024-01-08

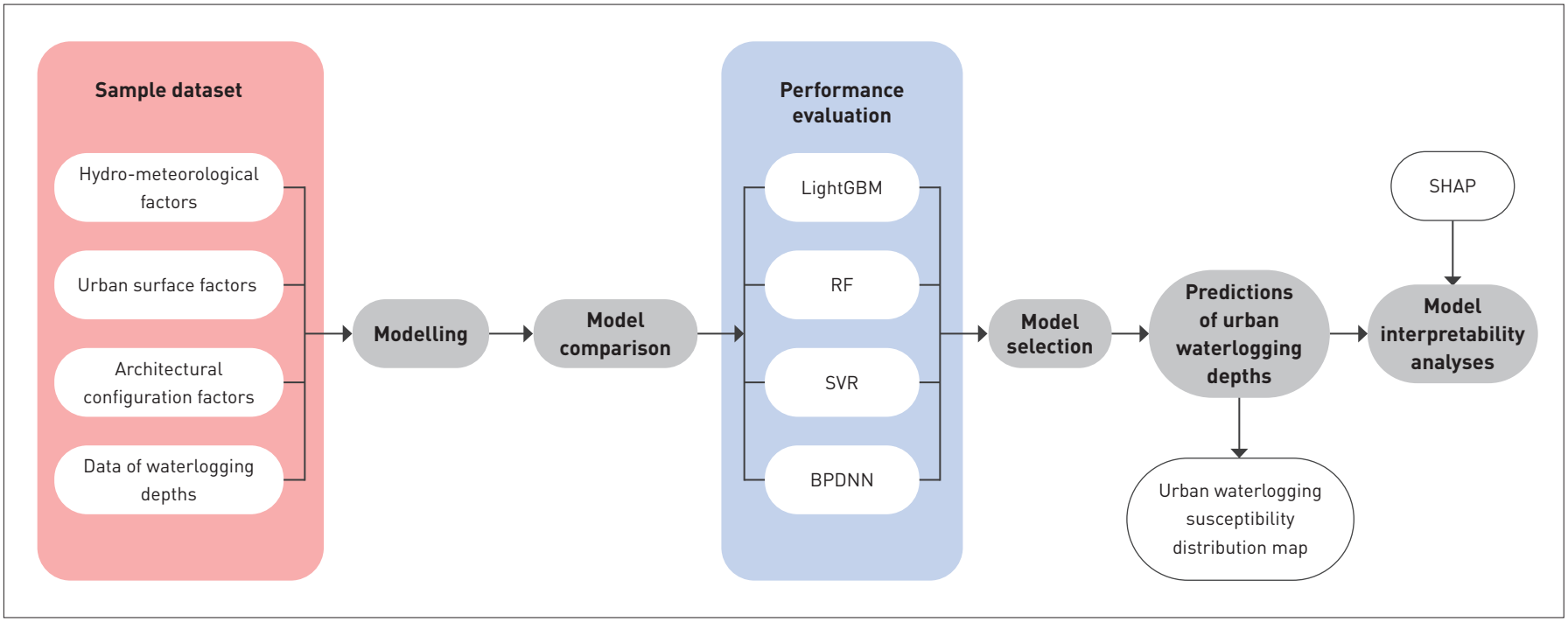
Prediction and Machine Learning Analysis of Urban Waterlogging Risks in High-Density Areas From the Perspective of the Built Environment: A Case Study of Shenzhen, China

Shiqi ZHOU[1], Weiyi JIA[2], Zhiyu LIU[3], Mo WANG[4,*]

1 College of Design and Innovation, Tongji University, Shanghai 200092, China
2 College of Architecture and Urban Planning, Tongji University, Shanghai 200092, China
3 Shanghai Tongji Urban Planning & Design Institute Co., Ltd., Shanghai 200082, China
4 College of Architecture and Urban Planning, Guangzhou University, Guangzhou 510006, China

***CORRESPONDING AUTHOR**
Address: No. 230, West Waihuan Road, College Town, Panyu District, Guangzhou 510006, Guangdong Province, China
Email: landwangmo@outlook.com

GRAPHICAL ABSTRACT



ABSTRACT

With the continuous advance of big data and artificial intelligence technologies, various data-driven machine learning algorithms have been widely applied in the studies of urban resilience, particularly in addressing the challenging issue of urban waterlogging. Currently, it is a pressing task to understand the influencing factors of waterlogging from the perspective of built environment, and provide guidance on dynamic monitoring and early alarm services. Focusing on Shenzhen, China, a typical high-density urbanized city, this research constructed a multifactorial dataset encompassing hydrological, meteorological, urban morphology, and waterlogging event data. Then, this research assessed and compared the performance of four mainstream machine learning models—LightGBM, RF, SVR, and BPDNN—in predicting urban waterlogging risks. The results showed that LightGBM had the best

accuracy and robustness in predicting waterlogging depths and risk levels in urban areas. The research also employed interpretability algorithm—Shapley Additive Explanations (SHAP)—for decoupling analysis. The results indicated that hydro-meteorological factors (the total rainfall volume and the rainfall lasting time) and several architectural configuration factors (e.g., density of buildings, building congestion degree) are the main influencing factors. In addition, the percentage of water body is vital to waterlogging regulation and retention, especially exhibiting a significant mitigating effect when exceeding 2.5%. This research provides a new technical method for urban waterlogging prediction and reveals the influencing factors and intrinsic mechanisms from the perspective of built environment, which is of great significance for the enhancement of the resilience of high-density cities.

KEYWORDS

Urban Waterlogging; Machine Learning; Model Performance Evaluation; Comparative Research; Model Interpretability Analysis; High-Density City

HIGHLIGHTS

- Proposes a comprehensive research framework combining LightGBM model and the interpretability algorithm of SHAP, and predicts waterlogging depth and its risk level in urban areas
- Verifies that historical downtowns in high-density cities face higher risks of waterlogging during extreme rainfall events with machine learning methods
- Presents a novel exploration in high-density urban context that analyzes the influencing factors and intrinsic mechanisms of urban waterlogging, focusing on hydro-meteorological, urban surface, and architectural configuration factor

RESEARCH FUND

"Resilience Enhancement and Dynamic Planning of Urban Grey-Green Infrastructure Based on Climate Adaptability," Guangdong Provincial Natural Science Foundation Youth Enhancement Project (No. 2023A1515030158)

EDITED BY Yuting GAO, Tina TIAN

1 Introduction

One of the core tasks in building urban resilience is accurately and effectively predicting urban risks and their impacts during spatial planning, alongside devising targeted adaptive planning strategies[1]. With the continuous advancement of artificial intelligence technologies, data-driven machine learning techniques have been widely applied in predicting urban waterlogging risks[2]~[5]. For instance, Elham Rafiei-Sardooi et al. utilized a Support Vector Machine model to map the flood vulnerability of the Khiyav Chai basin in Iran[3]; Zhaoli Wang et al. assessed the flood risk of the Dongjiang basin using a Random Forest model[4]. Compared with traditional hydrological and hydraulic models, the advantage of machine learning models lies in its ability to handle complex high-dimensional data with limited computational resources, especially in analyzing the nonlinear relationships between multivariate factors and target variables[5]. However, traditional machine learning models still face uncertainties in practice due to issues like overfitting limitation (e.g., only local optimum is supported when large datasets are used) and computational challenges (e.g., difficulty in generating optimal solutions with complex-structured data).

In recent years, a new generation of ensemble machine learning model, surpassing traditional algorithms in robustness, has emerged and been widely adopted in fields such as urban hydrological management[2][6]. The study by Hossein Shafizadeh-Moghadam et al. showed that ensemble machine learning models were more accurate and stable than traditional ones in flood susceptibility prediction[7]; Yuchen Guo et al. also confirmed that ensemble machine learning models significantly outperform the traditional model of Backpropagation Deep Neural Network in flood prediction[8]; Zening Wu et al. applied the ensemble machine learning model of Gradient Boosting Decision Tree (GBDT) to predict urban flood and waterlogging depths, validating its high accuracy[9]. Current research often focuses on the practicality of specific machine learning algorithms in urban flood prediction. However, there is less exploration of ensemble machine learning models in multi-scenario urban waterlogging prediction, resulting in a lack of detailed model comparisons and applications in spatial practice. This study aims to conduct a detailed comparative analysis of the ensemble algorithm LightGBM (Light Gradient Boosting Machine)[10] with three traditional machine learning algorithms, namely Random Forests (RF)[11], Support Vector Regression (SVR)[12], and Backpropagation Deep Neural Networks (BPDNN)[8], to reveal their performance differences in predicting

urban waterlogging risks in high-density areas and precisely dissect the influencing factors.

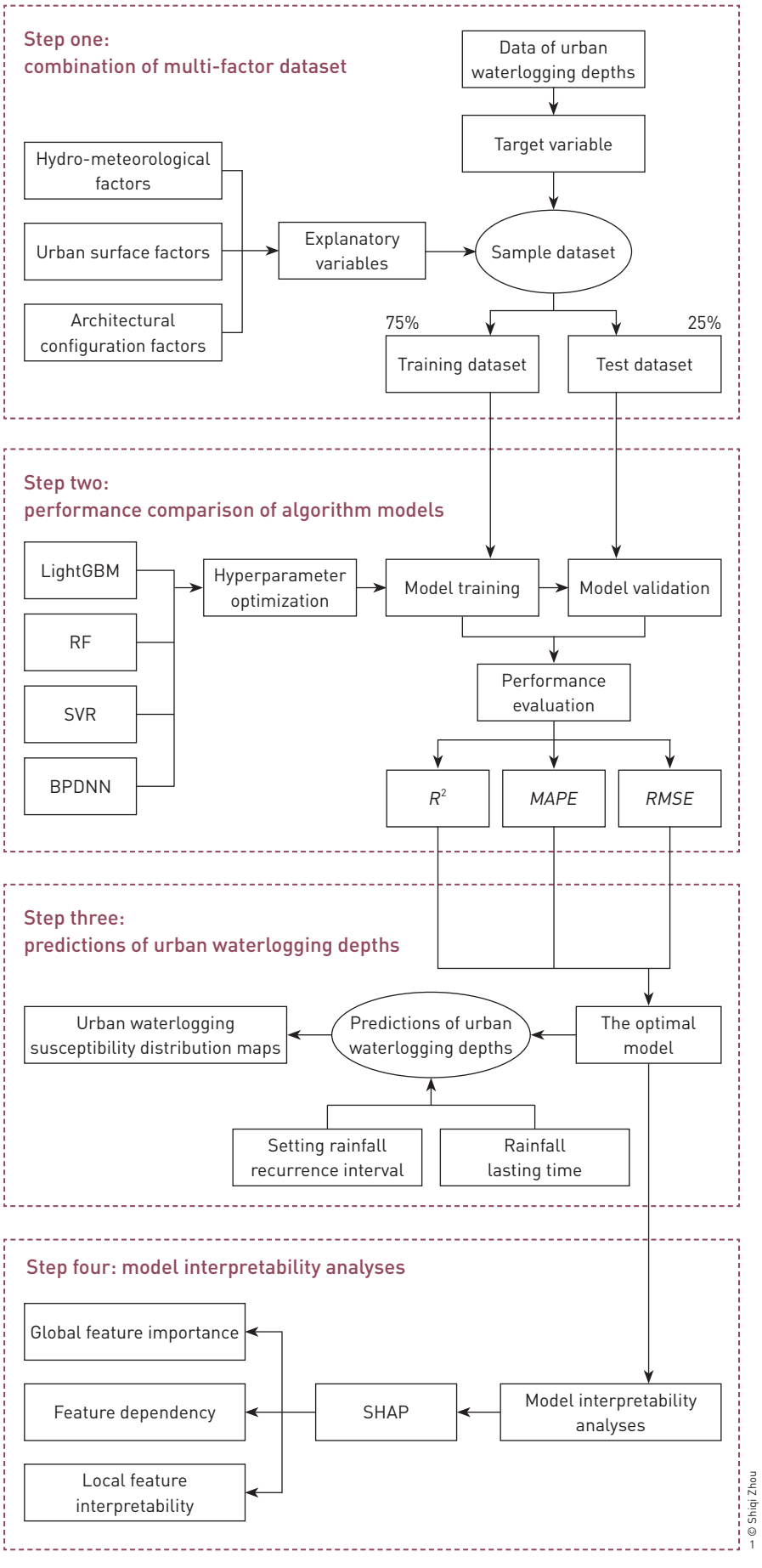
From the perspective of the built environment, this paper proposes a series of suggestions for enhancing urban resilience and provides an innovative method for predicting urban waterlogging risks in high-density areas, offering valuable insights and guidance for future urban planning theoretically and practically.

2 Study Area and Research Methods

This study focused on Shenzhen, a quintessential example of high-density urbanized city, as the case study, taking the city's precipitation events from January 1, 2019 to December 31, 2021 as samples. The research was initiated by constructing a multifactorial dataset for model training and testing, encompassing hydrological, meteorological, urban morphology, and waterlogging event data. The explanatory variables consisted of 3 categories (i.e., hydro-meteorological factors, urban surface factors, and architectural configuration factors) with 21 independent variables, while waterlogging depth being the target variable. The study assessed the performance of the four machine learning models—LightGBM, RF, SVR, and BPDNN—in predicting urban waterlogging risks. Based on the evaluation of model accuracy and robustness, this study selected the optimal model and generated the susceptibility distribution map of urban waterlogging risks in Shenzhen. Building on these predictions, the study then applied shapley additive explanations (SHAP) to conduct an in-depth analysis from global feature importance, feature dependency, and local feature interpretability, offering practical references for decision-making for urban resilience enhancement (Fig. 1).

2.1 Study Area

Nestled in the southern part of Guangdong Province along the eastern bank of the Pearl River Estuary, Shenzhen experiences a typical subtropical monsoon climate with prolonged summers, short winters, and copious precipitation. By the end of 2022, the city comprised 9 administrative districts and 1 functional district (Dapeng New District), spanning an area of 1,997.47 km^2[13] and homing a permanent population of approximately 17.66 million[14]. Studies have indicated that Shenzhen's urban waterlogging primarily results from the sudden and intense rainstorms in summer[15]. Statistics show that the city's annual rainfall averages 1,932.9 mm, of which approximately 86% occurs between April and September[16]. In recent years, along with the expansion of



1. Framework of model comparison and urban waterlogging risk assessment.

built-up areas, substantial alterations in urban spatial structure and surface conditions have progressively encroached upon urban green spaces and water retention areas, intensifying urban waterlogging risks.

2.2 Data Sources and Pre-processing

2.2.1 Data of Urban Waterlogging Depths

The data of urban waterlogging depths between 2019 and 2021 used in this study were sourced from 171 monitoring stations in Shenzhen, collected by curb membrane pressure sensors, with a sampling interval of 1 hour (Fig. 2). Taking into account the city's geographical and hydrological characteristics, the study area was divided into 171 sub-catchment units. The original sample data recorded all the rainfall events from January 1, 2019, to December 31, 2021, totaling 26,305 samples, which were further classified according to a 12-hour lag time (LTIME); then independent rainfall events were extracted from the continuous time series[17] and rainfall events with a total rainfall of less than 1 mm were excluded. After that, 167 samples of rainfall events which significantly impacted on urban waterlogging formed the final sample dataset.

2.2.2 Data of Impact Factors

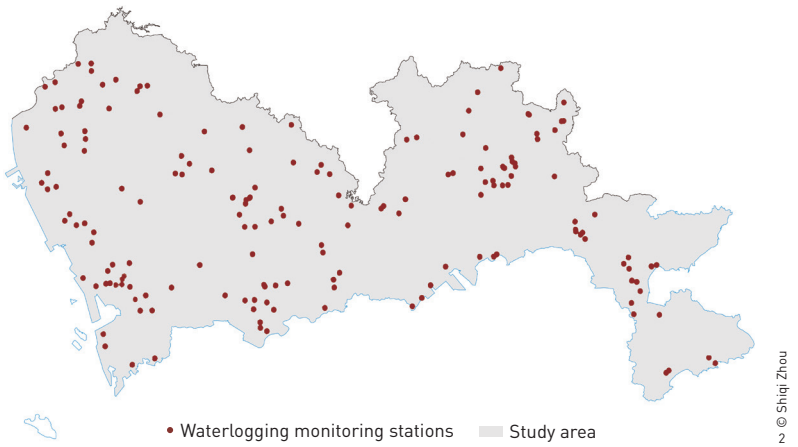
Existing literature has revealed that hydro-meteorological conditions[18]~[20], urban surface[4][21]~[24], and architectural configurations[4][25]~[27] primarily affect the occurrence and severity of urban waterlogging. Accordingly, this study selected 21 independent variables, comprising 2 hydro-meteorological factors[15][28], 10 urban surface factors[24][28]~[31] (Table 1), and 9 architectural configuration factors[25][29][32]~[34] (Table 2) as input features for the models. These factors were statistically analyzed for each sub-catchment unit.

2.2.3 Rainfall Scenario Simulation

As global climate change intensifies, cities are likely to experience extreme weather events more frequently in the future, exacerbating urban waterlogging risks. When assessing the waterlogging risk of a given area, it necessitates the simulation and prediction of multiple rainfall scenarios. This study established the one-hour LTIME and recurrence intervals of 1, 2, 3, 5, 10, and 20 years, thereby categorizing the 167 rainfall events into 6 scenarios. The storm intensity was calculated with Shenzhen's storm intensity formula[35]:

$$q = \frac{8.701(1 + 0.594)\lg R}{(t + 11.13)^{0.555}}, \quad (1)$$

where q represents the rainfall intensity, t denotes the LTIME, and R



2. Distribution map of 171 monitoring stations in Shenzhen (source: Meteorological Bureau of Shenzhen Municipality).

Table 1: Independent variables of hydro-meteorology and surface factors

Category	Independent variable	Source	Spatial Resolution (m)
Hydro-meteorology	Rainfall lasting time (LTIME)	Meteorological Bureau of Shenzhen Municipality	—
	Total rainfall volume (TOTAL_R)		
Surface	Percentage of green space (PGS)	GlobeLand 30	30 × 30
	Percentage of water body (PW)		
	Percentage of impervious surface (PIS)		
	Percentage of road (PR)		
	Average slope (AS)	Geospatial Data Cloud	30 × 30
	Average altitude (AA)		
	Average roughness (ARH)		
	Average relief (ARF)		
	Distance to river (DR)	Google Map	—
	Normalized difference vegetation index (NDVI)	NASA	250 × 250

Table 2: Architectural configuration factors

Independent Variable	Description	Unit	Formula	Source
Density of buildings (*DB*)	The number of buildings per unit area	n/hm^2	$DB = \frac{N}{A}$	OpenStreetMap
Mean building height (*MBH*)	The average height of buildings within a certain area	m	$MBH = \frac{\sum_{i=1}^{N} H_i}{N}$	
Mean building volume (*MBV*)	The average volume of buildings within a certain area	m^3	$MBV = \frac{\sum_{i=1}^{N} V_i}{N}$	
Standard deviation of building height (*SDBH*)	The average variation of building height within a certain area	m	$SDBH = \sqrt{\frac{\sum_{i=1}^{N}(H_i - MBH)^2}{N}}$	
Standard deviation of building volume (*SDBV*)	The average variation of building height within a certain area	m^3	$SDBV = \sqrt{\frac{\sum_{i=1}^{N}(V_i - MBV)^2}{N}}$	
Floor area ratio (*FAR*)	The ratio between the total building floor area and the total area of the site	—	$FAR = \frac{\sum_{i=1}^{N}(F_i \times S_i)}{A}$	
Building coverage ratio (*BCR*)	The ratio between the area covered by buildings and the total area of the site	—	$BCR = \frac{\sum_{i=1}^{N} S_i}{A}$	
Building shape coefficient (*BSC*)	The external skin surface divided by the volume of the buildings	m^{-1}	$BSC = \frac{\sum_{i=1}^{N} \frac{P_i \times H_i + S_i}{V_i}}{N}$	
Building congestion degree (*BCD*)	The total building volume divided by the highest building volume within a certain area	—	$BCD = \frac{\sum_{i=1}^{N} V_i}{\max(H_i) \times A}$	

NOTE

N represents the number of buildings within a specific area; *A* denotes the area of the given site; H_i refers to the height of building *i*; F_i indicates the number of floors of building *i*; S_i represents the area covered by building *i*; P_i is the perimeter of building *i*; V_i signifies the volume of building *i* (source: Refs. [25][34]).

stands for the recurrence interval. Chicago hyetograph method that most closely mirrors actual observational conditions[18][36] was then adopted to simulate the precipitation of different recurrence intervals.

2.3 Research Models and Methods

This study conducted hyperparameter optimization on 4 typical machine learning algorithm models. After training and testing, it utilized common model assessment metrics (R^2, *MAPE*, and *RMSE*) to identify the optimal model for predicting urban waterlogging depths in high-density areas. Additionally, the study used SHAP to analyze global feature importance, feature dependency, and local feature interpretability, so as to identify key factors influencing urban flooding and offer targeted recommendations for mitigation.

2.3.1 Models of Machine Learning Algorithms

(1) LightGBM

The ensemble algorithm LightGBM, developed by Microsoft, is a distributed gradient boosting algorithm built upon the Gradient Boosting Decision Tree (GBDT) framework[37], which stands as one of the most efficient machine learning algorithms currently available[10]. It introduces three major enhancements to the traditional GBDT algorithm structure: Gradient-based One-Side Sampling (GOSS), Exclusive Feature Bundling (EFB), and a histogram-based decision-making algorithm[38]. On the basis of the

given training dataset $\{(x_i, y_i)\}i=1^N$ (x_i represents the independent variables and y_i denotes the target variables), LightGBM aims to minimize the expected value of a specific loss function. The objective function is defined as follows:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \ , \tag{2}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \ , \tag{3}$$

where Obj represents the objective to be minimized, and $l(y_i, \hat{y}_i)$ is the loss function between the actual target value y_i and the predicted value $\hat{y}_i$. To prevent overfitting, a regularization term is defined as follows: the regularization component $\Omega(f_k)$ includes the complexity cost introduced by adding new leaf nodes, where f_k represents the tree model k, γ is the complexity cost of introducing new leaves, T is the number of leaf nodes in the tree, λ is the leaf weight adjustment coefficient, and ω denotes the weight value of the leaves.

(2) RF

RF, an ensemble learning model, was first introduced by Leo Breiman in 2001[11]. It is an enhanced version of the bagging decision tree, with the final classification or regression accomplished by a majority voting scheme among all individual decision trees[39], making it flexible and user-friendly.

(3) SVR

As a branch of SVM, SVR is primarily employed to address regression problems. Its objective is to minimize the distance between the hyperplane and the farthest sample points so that the data can be fitted by using the hyperplane. This process necessitates the selection of an appropriate kernel function. Studies have shown that the Radial Basis Function (RBF) outperforms other kernels in urban waterlogging prediction[12][40].

(4) BPDNN

BPDNN is a multilayer feedforward network trained via a backpropagation algorithm based on the error between the output and the desired output. BPDNN can handle nonlinear problems due to its multilayer structure, making it extensively used in the risk assessment of flood and waterlogging hazards.[8]

2.3.2 Assessment Methods of Model Performance

This study employed the metrics of R^2, *MAPE*, and *RMSE* to assess and compare the performance of the models. R^2 quantifies the fit of the regression model[41]; *MAPE* represents the relative error between the predictions and actual values[42]; *RMSE* evaluates the average error between predictions and actual values[43]. The mathematical formulas for these metrics are as follows, respectively:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \ , \tag{4}$$

$$MAPE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \ , \tag{5}$$

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \ , \tag{6}$$

where y_i refers to the observation i of waterlogging depth, $\hat{y}_i$ represents the corresponding predicted urban waterlogging depth (cm), and n denotes the number of observations.

2.3.3 SHAP Algorithm

The challenge of interpretability, commonly known as the "black box" issue, remains one of the drawbacks of ensemble machine learning[44]. The SHAP algorithm, based on cooperative game theory, constructs an additive explanation model that estimates the contribution of each feature. Numerous recent studies have started employing SHAP to visually explain complex ensemble models, validating its strong credibility[45][46]. This study adopted the SHAP algorithm to investigate the global relationships and importance of various features with respect to urban waterlogging depths, while also conducting local feature interpretability analysis, as follows:

$$g(z') = \phi_0 + \sum_{i=1}^{M}\phi_i z_i' \ , \tag{7}$$

where g represents the explanation model, $z' \in \{0, 1\}^M$ is the coalition vector, M is the maximum coalition size, and ϕ_i denotes the contribution of feature i to the model output.

In the SHAP algorithm, the contribution of each feature is determined by their marginal contributions to the output[47]. This facilitates the interpretation of the machine learning model from both global and local perspectives. Shapley values not only reflect the importance of each feature but also indicate the positive or negative impact on the target variable[48].

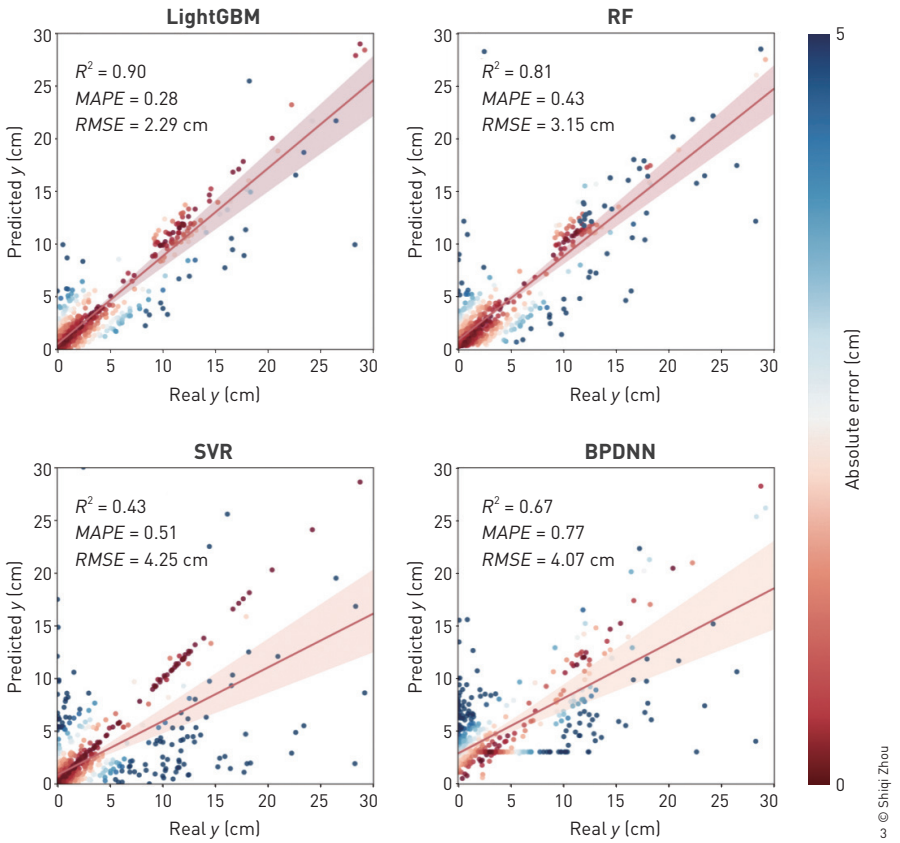
3 Results and Discussion

3.1 Performance Comparison of Algorithm Models

The evaluation of model performance on the training dataset (Table 3) indicated that LightGBM (R^2 = 0.96) had a significantly better fit than RF (R^2 = 0.89), SVR (R^2 = 0.78), and BPDNN (R^2 = 0.69). Similarly, LightGBM exhibited the highest fitting accuracy (*MAPE* = 0.22), with the smallest relative error, exceeded by RF (*MAPE* = 0.32), SVR (*MAPE* = 0.44), and BPDNN (*MAPE* = 0.63). As for *RMSE*, LightGBM showed the lowest value (*RMSE* = 1.81 cm), indicating minimal deviation and the best predictive accuracy, followed by RF (*RMSE* = 2.26 cm), SVR (*RMSE* = 3.12 cm), and BPDNN (*RMSE* = 3.64 cm). Overall, the sound precision and stability of LightGBM was demonstrated.

To verify the robustness of the models in predicting urban waterlogging depth, 25% of the rainfall events (42, totaling 1,790 samples) were used as an independent test dataset based on the algorithm logics of machine learning. The results confirmed that LightGBM consistently outperformed other models in predicting urban waterlogging depth on the test dataset (R^2 = 0.90, *MAPE* = 0.28, and *RMSE* = 2.29 cm).

The results showed that the best-fit lines of LightGBM and RF are closer to the ideal-fit line of "$y = x$" (Fig. 3). Further comparative analysis revealed that although RF performed comparably with LightGBM when urban waterlogging depths were less than 10 cm, LightGBM excelled when the depths grew. Considering the comprehensive metric evaluation and the distribution of absolute error points, it can be concluded that LightGBM significantly surpassed the other three machine learning models in terms of prediction accuracy and robustness. Thus, this study employed LightGBM to predict urban waterlogging depths under different rainfall scenarios in Shenzhen and create corresponding urban waterlogging risk maps.



3. Scatter plot and fitting lines based on the predicted and observed urban waterlogging depths of the test dataset (the red line represents the best-fit line, and the red area indicates the 95% confidence interval).

Table 3: Performance evaluation of 4 models on training dataset and test dataset

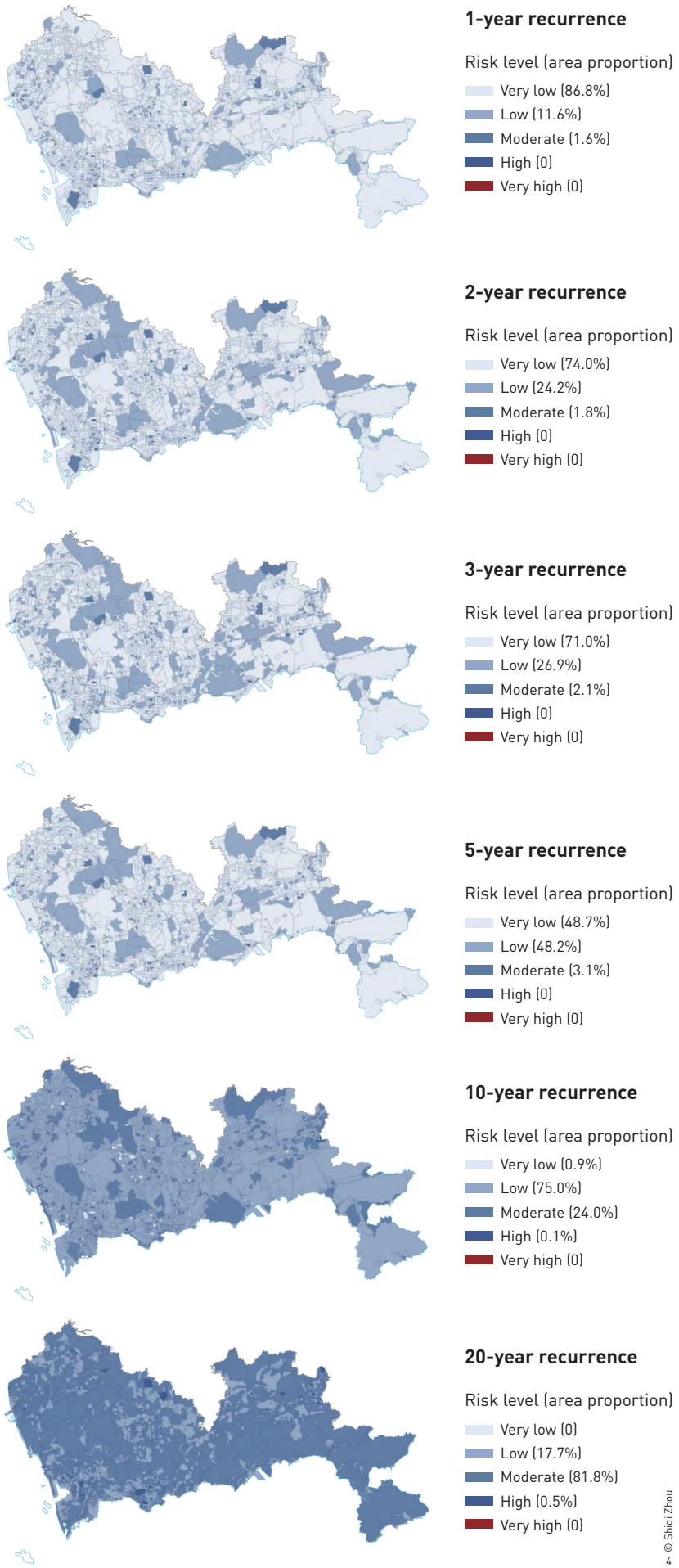
Model	R^2		*MAPE*		*RMSE* (cm)	
	Training	Test	Training	Test	Training	Test
LightGBM	0.96	0.90	0.22	0.28	1.81	2.29
RF	0.89	0.81	0.32	0.43	2.26	3.15
SVR	0.78	0.43	0.44	0.51	3.12	4.25
BPDNN	0.69	0.67	0.63	0.77	3.64	4.07

NOTE
When an R^2 value is closer to 1, it indicates a higher degree of model fitting; when a *MAPE* value is closer to 0, it indicates a greater accuracy of the predictions; when an *RMSE* value is closer to 0, it signifies a smaller prediction error and a stronger predictive capability of the model.

3.2 Predictions of Urban Waterlogging Depths

Utilizing the natural break method, this study classified the peak waterlogging depth risks in the sub-catchment units into 5 categories: very low (0 ~ 5 cm), low (5 ~ 10 cm), moderate (10 ~ 15 cm), high (15 ~ 20 cm), and very high (≥ 20 cm). The study set the one-hour duration and compared the waterlogging risk variations across different recurrence intervals—1, 2, 3, 5, 10, and 20 years (Fig. 4).

As the rainfall recurrence interval increased from 1 year to 10, the proportion of very low risk zones in Shenzhen gradually decreased, while those considered low and medium risk zones correspondingly increased (Fig. 4). During this phase, no high and very high risk zones were observed, indicating that despite local waterlogging may



4. Urban waterlogging susceptibility distribution maps under different rainfall recurrence intervals for one-hour duration.

occur in the areas with outdated drainage facilities, the majority of Shenzhen's drainage systems can effectively respond to moderate rainfall events. However, when the recurrence interval extended to 20 years, the spatial distribution of urban waterlogging risk significantly changed, particularly with a notable increase of medium risk zones. This revealed that most areas in Shenzhen would experience varying degrees of urban waterlogging during 20-year rainfall events, with high risk zones mainly found in older districts with complex drainage and terrain conditions, such as Nanshan, Futian, and Pingshan Districts.

3.3 Model Interpretability Analyses

3.3.1 Global Feature Importance

Global feature importance quantifies the impact of each feature on prediction outcome within a given model, primarily assessed through the absolute average of each variable's Shapley value[49]. By calculating all the features (Fig. 5), the analysis revealed that among the top 60% of the features, the hydro-meteorological factors, especially TOTAL_R and LTIME, had a profoundly significant impact on the model, each contributed 24.6% and 15.8% of the importance; PW and PIS had importance contributions of 6.4% and 5.1%, and *BCD* and *DB* contributed 4.4% and 4.1% of the importance, respectively.

Overall, hydro-meteorological factors significantly outweighed the impact of the other two kinds of factors on urban waterlogging depth, confirming previous studies that identified rainfall amount and duration as the main inducing factors of urban waterlogging[18]~[20]. Meanwhile, PW was the only feature that had a mitigating effect on urban waterlogging. This somehow explains why many cities construct artificial lakes within urban areas—to capture runoffs during extreme rainfall events, alleviating pressure on downstream water bodies and stormwater infrastructure networks.

Among the architectural configuration factors, both *BCD* and *DB* describe building density. As these two features increase, there is a corresponding rise in PIS, which leads to a reduction in the stormwater retention capacity of blue-green infrastructure, thereby increasing the likelihood of urban waterlogging.

3.3.2 Feature Dependency

To delve into the nonlinear relationships between urban waterlogging depth and the primary disaster-inducing factors, this

study employed feature dependency plots of SHAP to reveal the contribution of individual variables to a given model's output. The feature dependency plot illustrates the extent to which a particular feature modifies the model's prediction, visualizing the marginal effects between feature values and their corresponding Shapley values. The sign of Shapley values (+/−) indicates whether a specific feature has a positive or negative impact on the prediction[50]. According to the results of the global importance analysis, the six most highly ranked features from the three categories of factors were selected for further analysis.

(1) Hydro-meteorological factors

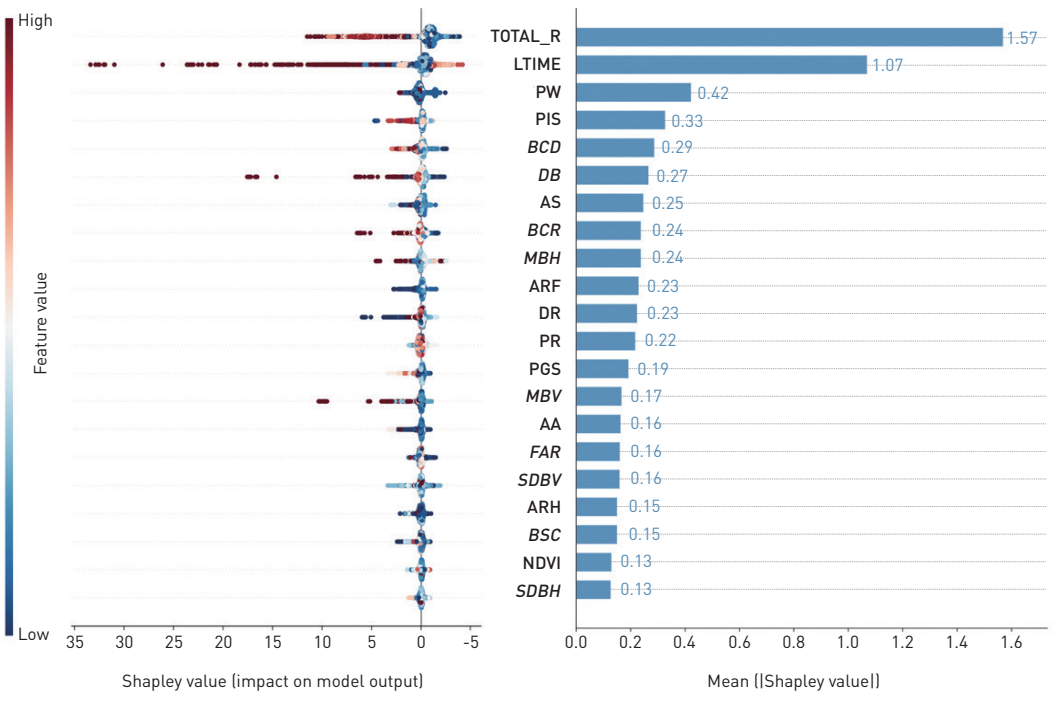
The results demonstrated a generally positive correlation between TOTAL_R and urban waterlogging depth (Fig. 6-1). When TOTAL_R was below 25 mm, it slightly impacted on urban waterlogging. However, as TOTAL_R exceeded 25 mm, its impact manifested by three stages: initially, urban waterlogging risk escalated as the total rainfall increased from 25 mm to 100 mm, with the average Shapley value rising from 2 to approximately 6; the impact of rainfall stabilized when TOTAL_R was between 100 mm and 125 mm, indicating a marginal effect that additional increases in rainfall do not exacerbate urban waterlogging; when the rainfall surpassed 125 mm, the average Shapley value climbed swiftly again, and urban waterlogging risk kept increasing, possibly due to the limitation of the urban drainage system's capacity. Although local variations in urban environment and infrastructure might affect their sensitivity to rainfall, in general, the strong influence of rainfall on urban waterlogging risk is consistent across nearly all regions.

The dependency plot of Shapley value for LTIME (Fig. 6-2) showed that when the duration of continuous rainfall was less than 55 hours, LTIME had a fluctuating impact on the model's output. Although prolonged LTIME exacerbated urban waterlogging risk, it did not exhibit a linear growth effect similar to that of TOTAL_R. However, once the duration exceeded 55 hours, the Shapley values surged dramatically. The study suggests that for LTIME less than 55 hours, the hourly rainfall of most events did not reach the disaster-inducing threshold, thus exerting a less intense pressure on urban waterlogging depth.

(2) Urban surface factors

The feature dependency plot for urban surface factors indicated that PW can lower the urban waterlogging risks (Fig. 7-1): the higher proportion of water bodies correlated with stronger mitigation of urban waterlogging; after reaching a certain threshold (12.5%), it exhibited no further impacts on waterlogging. When PW was below 1.2%, the stormwater retention capacity of water surfaces was minimal. However, once the proportion exceeded 2.5%, the mitigation effect on urban waterlogging gradually augmented. According to the research by Wenchao Qi et al., urban lakes can provide substantial buffer areas for runoff during flood seasons, thereby facilitating urban waterlogging management[51]. Hence, preserving natural lakes and strategically constructing urban artificial lakes are beneficial for stormwater management.

The impact of PIS on urban waterlogging can be unfolded in

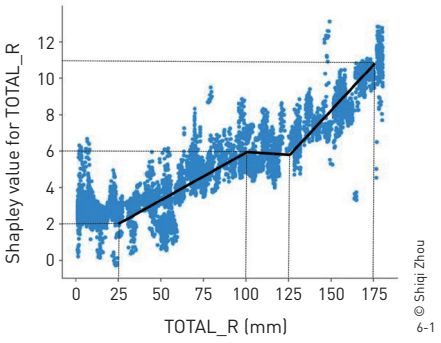


5. Scatter plot of feature Shapley values (left) and global feature importance ranking plot (right) of LightGBM model.
6. Dependency plot of important variables and Shapley value (the black line marks the division between different stages): dependency plot of Shapley value for TOTAL_R (Fig. 6-1); dependency plot of Shapley value for LTIME (Fig. 6-2).
7. Dependency plot of important variables and Shapley value: dependency plot of Shapley value for PW (Fig. 7-1); dependency plot of Shapley value for PIS (Fig. 7-2).
8. Dependency plot of important variables and Shapley value: dependency plot of Shapley value for *BCD* (Fig. 8-1); dependency plot of Shapley value for *DB* (Fig. 8-2).

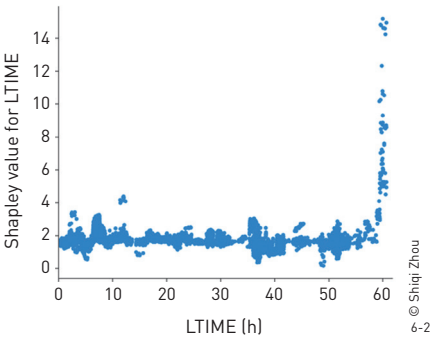
three stages (Fig. 7-2): 1) shifting from 0 to 15%, the Shapley values remained close to zero, suggesting a slight impact; 2) the impact gradually intensified once exceeding the 15% threshold and peaked at 30%①; and 3) there observed no further rise if beyond 30%. Relevant research indicated that the spatial configuration of impervious surfaces significantly affects urban waterlogging. Reducing PIS can provide sufficient space for vegetation to intercept and for soil to absorb rainwater, thereby lowering peak runoff and mitigating urban waterlogging[52].

(3) Architectural configuration factors

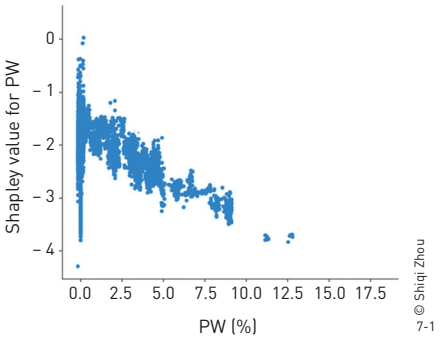
In terms of architectural configuration, *BCD* had an impact on urban waterlogging depth (Fig. 8-1), with a noticeable increase in average Shapley value within [0.01, 0.02]. Examination of sample data revealed that these sub-catchment units were primarily covered by hard paving; despite their low *DB*, the extensive impervious surfaces impeded effective infiltration and retention of rainfall, leading to an increase of transient runoff and thus triggering urban waterlogging. Therefore, the proportion of impervious pavement should be strictly controlled in urban design or renewal practice. A more significant compressive effect on urban waterlogging was observed when *BCD* exceeded the threshold of 0.08. Meanwhile, *DB* exhibited clear spatial heterogeneity in its impact on urban waterlogging (Fig. 8-2). Even with the same *DB* conditions, differences in building layout, height, and morphology can lead to variations of Shapley values. When *DB* is below 15 buildings per hectare, Shapley values fluctuated between [– 0.5, 1.0]; while above this *DB* threshold, Shapley values ranged [0.0, 2.0], indicating a significant increase of urban waterlogging risk.



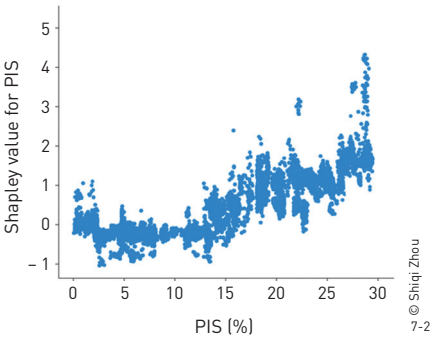
6-1



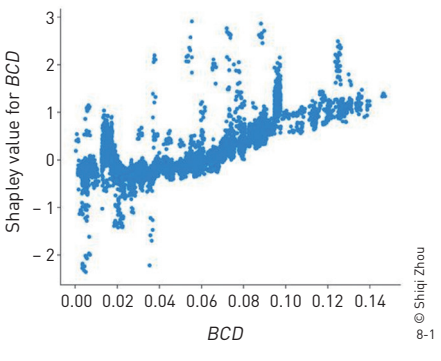
6-2



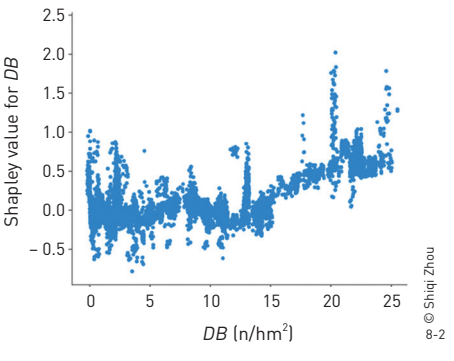
7-1



7-2



8-1



8-2

3.3.3 Local Feature Interpretability

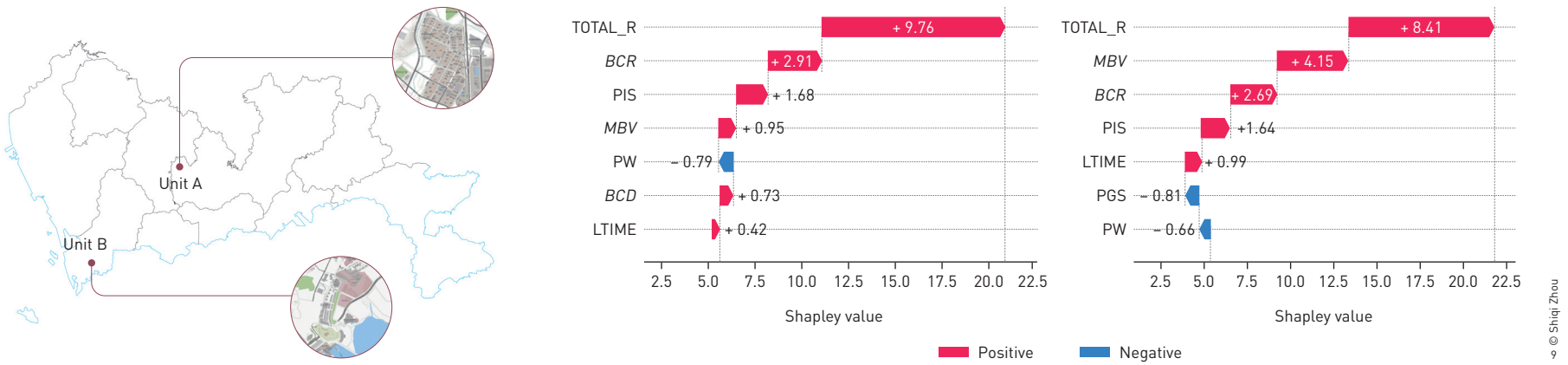
The contribution analysis of different features to the model output for each sub-catchment units can provide an in-depth understanding of the impact of spatial heterogeneity on urban waterlogging, and inform the development of adaptive strategies. Based on previous predictions (Fig. 4), two sub-catchment units of high waterlogging risk in Shenzhen were selected for individual sample feature analysis, where unit A represents a high-density historical downtown of Shenzhen and unit B represents a newly developed coastal area of the city (Fig. 9)②.

For unit A, the features of TOTAL_R, *BCR*, PIS, and *MBV* accounted for the top 80% of the contribution. Consistent with the results of previous global feature importance analysis, TOTAL_R was the most significant factor affecting urban waterlogging risk in the unit; the other three were architectural configuration factors. Locating in the old city area, unit A had an extreme high *DB* with complex road system and a large coverage of impervious surfaces, aggravating substantial runoff during extreme rainfall events.

In unit B, the features contributing to the top 80% of the importance were TOTAL_R, *MBV*, *BCR*, PIS, and LTIME. Additionally, possibly due to the differences in PGS, comparing to unit A, the contribution of PGS obviously increased. This could be attributed to the location of unit B in the coastal commercial and high-end residential areas, where, despite the densely distributed buildings, coastal greenbelts, wetland parks, and roadside green space have

① When PIS exceeded 30%, the impacts on urban waterlogging were not strengthened. Thus it mainly presents the range of 0 to 30% in Fig. 7.

② Since the contribution value of the features at a lower rank is too small and not typical for further discussion, the local feature interpretability analysis focused on the top 7 contributing features.



9. Sample locations of unit A and unit B (left), and the rankings of local feature contributions (right).

together highlighted the contribution of PGS in the model, exerting a mitigating effect on urban waterlogging.

4 Conclusions and Perspectives

Based on the observed data of urban waterlogging records in Shenzhen between 2019 and 2021, this study integrated hydro-meteorological, urban surface, and architectural configuration factors to predict urban waterlogging risk using four machine learning models—LightGBM, RF, SVR, and BPDNN. By comparing the performance of the models, LightGBM was selected as the optimal predictive model for urban waterlogging risk assessment, leading to the following conclusions.

When experiencing high-recurrence rainfall events (e.g., once every 20 years), the high risk zones in Shenzhen will emerge predominantly in the old city districts of Nanshan, Futian, and Pingshan. Feature analysis of the models indicated that hydro-meteorological factors (including TOTAL_R and LTIME) were the primary disaster-inducing elements, contributing 40.4% to the model. The impact on urban waterlogging became particularly pronounced when TOTAL_R exceeded 125 mm or LTIME exceeds 55 hours. PW was the only feature showing a mitigating effect on urban waterlogging, and when it exceeds 2.5% the stormwater regulation and retention capacity enhanced. Features reflecting building density exhibited significantly positive correlation with urban waterlogging risk once exceeding certain thresholds. Notably, the interpretability of urban surface factors like ARF was relatively low, which may be related to the minor topographical changes and high imperviousness in Shenzhen's built-up areas. TOTAL_R and *DB* remained the primary features locally affecting urban waterlogging depth, while in certain regions, blue-green infrastructure played a crucial role in mitigating urban waterlogging.

The LightGBM-based method for predicting urban waterlogging risk proposed and validated in this study is of universal significance. The analysis of critical factors affecting urban waterlogging through interpretability algorithms can provide guidance for urban planning and construction. In the development of high-density urban areas, it is necessary to strengthen the renovation of old neighborhoods, restore natural ecosystems, promptly improve the capacities of drainage and flood prevention infrastructure. It also needs to restore and expand water bodies for natural regulation space expansion in and around cities, and to construct flood storage and security engineering projects according to relevant standards and plans. Furthermore, it is necessary to create more open green spaces integrating spatial and vertical design in urban construction and renewal; adaptively increase the proportion of permeable pavement; and promote the construction of sponge cities to preserve natural rainwater and flood channels and storage spaces including rivers, lakes, and wetlands, establishing a comprehensive ecological infrastructure system.

Due to the constraints in acquiring urban pipeline network data[28], this study employed road factor (PR) as a proxy. This can reflect the efficiency of urban drainage to a certain degree, but there are still some limitations. Utilizing data with higher spatial resolution would reveal urban waterlogging dynamics with greater details and increase the precision of the identification and analysis of the key factors. Future studies could also integrate hydrological and hydraulic models for more targeted and accurate experiments.

Competing interests | The authors declare that they have no competing interests.

REFERENCES

[1] Jha, A. K., Miner, T. W., & Stanton-Geddes, Z. (Eds.). (2013). *Building Urban Resilience: Principles, Tools, and Practice*. The World Bank.

[2] Arabameri, A., Saha, S., Chen, W., Roy, J., Pradhan, B., & Bui, D. T. (2020). Flash flood susceptibility modelling using functional tree and hybrid ensemble techniques. *Journal of Hydrology*, (587), 125007.

[3] Rafiei-Sardooi, E., Azareh, A., Choubin, B., Mosavi, A. H., & Clague, J. J. (2021). Evaluating urban flood risk using hybrid method of TOPSIS and machine learning. *International Journal of Disaster Risk Reduction*, (66), 102614.

[4] Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, (527), 1130–1141.

[5] Chen, J., Huang, G., & Chen, W. (2021). Towards better flood risk management: Assessing flood risk and investigating the potential mechanism based on machine learning models. *Journal of Environmental Management*, (293), 112810.

[6] Mei, C., Liu, J., Wang, H., Yang, Z., Ding, X., & Shao, W. (2018). Integrated assessments of green infrastructure for flood mitigation to support robust decision-making for sponge city construction in an urbanized watershed. *Science of the Total Environment,* (639), 1394–1407.

[7] Shafizadeh-Moghadam, H., Valavi, R., Shahabi, H., Chapi, K., & Shirzadi, A. (2018). Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. *Journal of Environmental Management*, (217), 1–11.

[8] Guo, Y., Quan, L., Song, L., & Liang, H. (2022). Construction of rapid early warning and comprehensive analysis models for urban waterlogging based on AutoML and comparison of the other three machine learning algorithms. *Journal of Hydrology*, (605), 127367.

[9] Wu, Z., Zhou, Y., Wang, H., & Jiang, Z. (2020). Depth prediction of urban flood under different rainfall return periods based on deep learning and data warehouse. *Science of the Total Environment*, (716), 137077.

[10] Gan, M., Pan, S., Chen, Y., Cheng, C., Pan, H., & Zhu, X. (2021). Application of the machine learning LightGBM model to the prediction of the water levels of the lower Columbia River. *Journal of Marine Science and Engineering*, 9(5), 496.

[11] Breiman, L. (2001). Random forests. *Machine Learning*, (45), 5–32.

[12] Panahi, M., Dodangeh, E., Rezaie, F., Khosravi, K., Van Le, H., Lee, M.-J., Lee, S., & Pham, B. T. (2021). Flood spatial prediction modeling using a hybrid of meta-optimization and support vector regression modeling. *CATENA*, (199), 105114.

[13] Community Construction and Zoning Office, Bureau of Civil Affairs of Shenzhen Municipality. (2024, April 3). *Overview of administrative division information*.

[14] Statistics Bureau of Shenzhen Municipality. (2023, May 8). *Shenzhen statistical bulletin on 2022 national economic and social development*.

[15] Zhou, S., Liu, Z., Wang, M., Gan, W., Zhao, Z., & Wu, Z. (2022). Impacts of building configurations on urban stormwater management at a block scale using XGBoost. *Sustainable Cities and Society*, (87), 104235.

[16] Meteorological Bureau of Shenzhen Municipality. (2024, May 15). *Climatic profile and seasonal characteristics of Shenzhen*.

[17] Ke, Q., Tian, X., Bricker, J., Tian, Z., Guan, G., Cai, H., Huang, X., Yang, H., & Liu, J. (2020). Urban pluvial flooding prediction by machine learning approaches—A case study of Shenzhen City, China. *Advances in Water Resources*, (145), 103719.

[18] Hou, J., Guo, K., Wang, Z., Jing, H., & Li, D. (2017). Numerical simulation of design storm pattern effects on urban flood inundation. *Advances in Water Science*, 28(6), 820–828.

[19] Zhou, H., Liu, J., Gao, C., & Ou, S. (2018). Analysis of current situation and problems of urban waterlogging control in China. *Journal of Catastrophology*, 33(3), 147–151.

[20] Song, L., & Xu, Z. (2019). Coupled hydrologic-hydrodynamic model for urban rainstorm water logging simulation: Recent advances. *Journal of Beijing Normal University (Natural Science)*, 55(5), 581–587.

[21] Wu, J., & Zhang, P. (2017). The effect of urban landscape pattern on urban waterlogging. *Acta Geographica Sinica*, 72(3), 444–456.

[22] Xu, Y., Li, K., Xie, Y., Ling, H., Qian, M., Wang, X., & Lu, Y. (2018). Study on the influencing factors and multiple regression model of urban waterlogging based on GIS—A case study of Shanghai. *Journal of Fudan University (Natural Science)*, 57(2), 182–198.

[23] Xu, H., Lu, H., Zhan, X., Li, J., Gao, C., & Zhang, T. (2024). Impacts of underlying surface changes and rainfall patterns on flooding at airport area in Zhuhai. *China Rural Water and Hydropower*, 1–16.

[24] Shrestha, R., Di, L., Eugene, G. Y., Kang, L., Shao, Y.-Z., & Bai, Y.-Q. (2017). Regression model to estimate flood impact on corn yield using MODIS NDVI and USDA cropland data layer. *Journal of Integrative Agriculture*, 16(2), 398–407.

[25] Lin, J., He, X., Lu, S., Liu, D., & He, P. (2021). Investigating the influence of three-dimensional building configuration on urban pluvial flooding using random forest algorithm. *Environmental Research*, (196), 110438.

[26] Kim, Y., Eisenberg, D. A., Bondank, E. N., Chester, M. V., Mascaro, G., & Underwood, B. S. (2017). Fail-safe and safe-to-fail adaptation: Decision-making for urban flooding under climate change. *Climatic Change*, (145), 397–412.

[27] Wang, J., Yu, C. W., & Cao, S.-J. (2022). Urban development in the context of extreme flooding events. *Indoor and Built Environment*, 31(1), 3–6.

[28] Wang, M., Li, Y., Yuan, H., Zhou, S., Wang, Y., Ikram, R. M. A., & Li, J. (2023). An XGBoost-SHAP approach to quantifying morphological impact on urban flooding susceptibility. *Ecological Indicators*, (156), 111137.

[29] Yan, M., Yang, J., Ni, X., Liu, K., Wang, Y., & Xu, F. (2024). Urban waterlogging susceptibility assessment based on hybrid ensemble machine learning models: A case study in the metropolitan area in Beijing, China. *Journal of Hydrology*, (630), 130695.

[30] Zhang, H., Zhang, J., Fang, H., & Yang, F. (2022). Urban flooding response to rainstorm scenarios under different return period types. *Sustainable Cities and Society*, (87), 104184.

[31] Kumar, R., & Acharya, P. (2016). Flood hazard and risk assessment of 2014 floods in Kashmir Valley: A space-based multisensor approach.

Natural Hazards, (84), 437–464.

[32] Jiang, F., Xie, Z., Xu, J., Yang, S., Zheng, D., Liang, Y., Hou, Z., & Wang, J. (2023). Spatial and component analysis of urban flood resiliency of Kunming City in China. *International Journal of Disaster Risk Reduction*, (93), 103759.

[33] Xu, Y., Liu, M., Hu, Y., Li, C., & Xiong, Z. (2019). Analysis of three-dimensional space expansion characteristics in old industrial area renewal using GIS and Barista: A case study of Tiexi District, Shenyang, China. *Sustainability*, 11(7), 1860.

[34] Cheng, C., Yu, X., Guo, S., & Ma, T. (2005). Analysis of the crowd degree of building for communities based on high spatial resolution remote sensed images. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 41(6), 875–881.

[35] Meteorological Bureau of Shenzhen Municipality. (2023, March 30). *New version of the storm intensity formula*.

[36] Dai, Y., Wang, Z., Dai, L., Cao, Q., & Wang, T. (2017). Application of Chicago Hyetograph Method in design of short duration rainstorm patterns. *Journal of Arid Meteorology*, 35(6), 1061–1069.

[37] Wu, Z., Qiao, R., Zhao, S., Liu, X., Gao, S., Liu, Z., Ao, X., Zhou, S., Wang, Z., & Jiang, Q. (2022). Nonlinear forces in urban thermal environment using Bayesian optimization-based ensemble learning. *Science of the Total Environment*, (838), 156348.

[38] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*. Neural Information Processing Systems Foundation, Inc.

[39] Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, (114), 24–31.

[40] Wu, J., Liu, H., Wei, G., Song, T., Zhang, C., & Zhou, H. (2019). Flash flood forecasting using support vector regression model in a small mountainous catchment. *Water*, 11(7), 1327.

[41] Wu, J., Liu, Z., Liu, T., Liu, W., Liu, W., & Luo, H. (2023). Assessing urban pluvial waterlogging resilience based on sewer congestion risk and climate change impacts. *Journal of Hydrology*, (626), 130230.

[42] Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, 15(4), 405–408.

[43] Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, 15(14), 5481–5487.

[44] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2023). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, (16), 45–74.

[45] Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers, Environment and Urban Systems*, (96), 101845.

[46] Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention*, (136), 105405.

[47] Van den Broeck, G., Lykov, A., Schleich, M., & Suciu, D. (2022). On the tractability of SHAP explanations. *Journal of Artificial Intelligence Research*, (74), 851–886.

[48] Molnar, C. (2020). *Interpretable machine learning*.

[49] Casalicchio, G., Molnar, C., & Bischl, B. (2019). Visualizing the Feature Importance for Black Box Models. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I* (pp. 655–670). Springer.

[50] Michiels, J., Suykens, J., & De Vos, M. (2024). Explaining the model and feature dependencies by decomposition of the Shapley value. *Decision Support Systems*, (182), 114234.

[51] Qi, W., Hou, J., Liu, J., Han, H., Guo, K., & Ma, Y. (2018). Lake control on surface runoff causing urban flood inundation. *Journal of Hydroelectric Engineering*, 37(9), 8–18.

[52] Poelmans, L., Van Rompaey, A., & Batelaan, O. (2010). Coupling urban expansion models and hydrological models: How important are spatial patterns?. *Land Use Policy*, 27(3), 965–975.

https://doi.org/10.15302/J-LAF-0-020023 收稿日期 | 2023-09-03 录用日期 | 2024-01-08 中图分类号 | TU984.11+6, TP181 文献标识码 | A

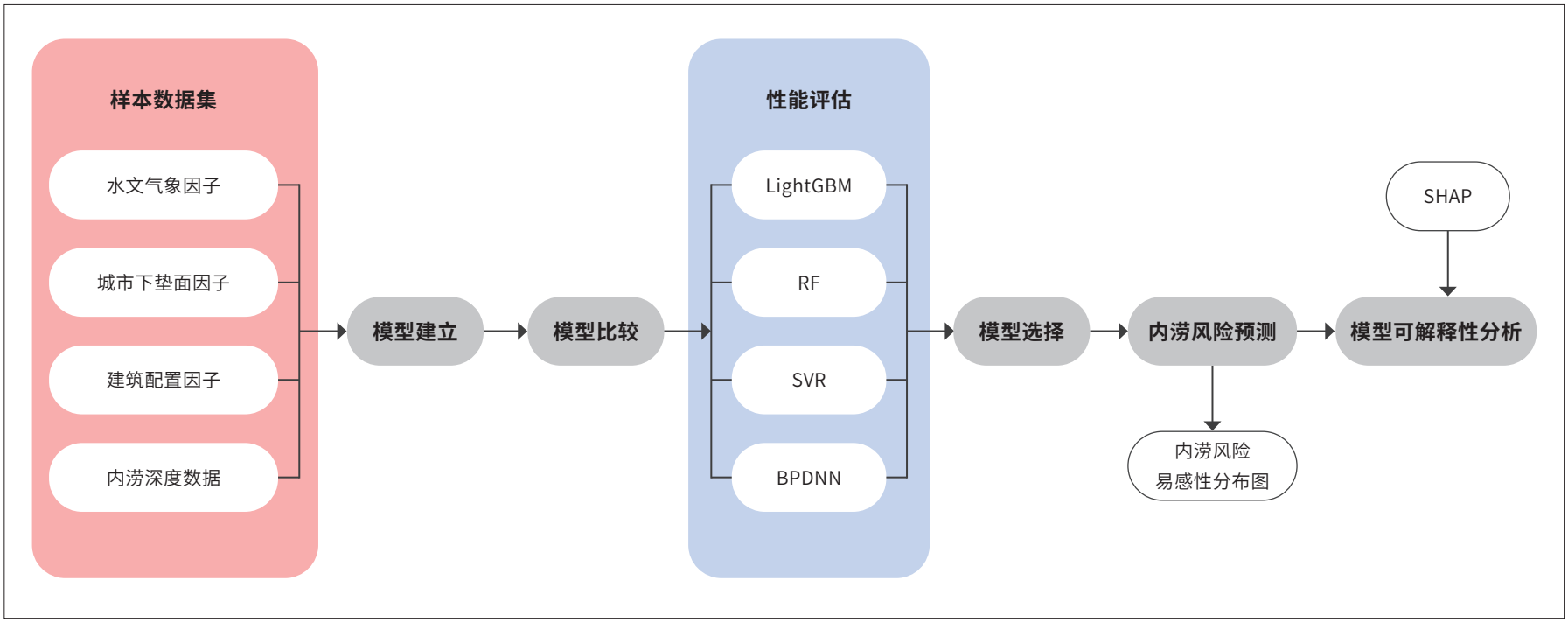
建成环境视角下高密度城市内涝风险预测与影响因素机器学习解析：以深圳市为例

周士奇[1]，贾蔚怡[2]，刘治宇[3]，王墨[4,*]

1 同济大学设计创意学院，上海 200092
2 同济大学建筑与城市规划学院，上海 200092
3 上海同济城市规划设计研究院有限公司，上海 200082
4 广州大学建筑与城市规划学院，广州 510006

***通讯作者**
地址：广东省广州市番禺区大学城外环西路230号
邮编：510006
邮箱：landwangmo@outlook.com

图文摘要



摘要

伴随着大数据和人工智能技术的演进，多种基于数据驱动的机器学习算法已逐渐在城市韧性研究中得到广泛应用，尤其是针对城市内涝这一关键问题。目前，解决城市内涝的重要任务是从建成环境的角度理解内涝影响因素，并指导动态监测和预警服务。本研究选取深圳市作为高密度城市的典型代表，构建了涵盖水文、气象、城市形态和内涝事件等多方面数据的多因子联动数据集，并对比了四种主流机器学习模型（LightGBM、RF、SVR和BPDNN）在预测城市内涝风险方面的性能差异。结果显示，LightGBM在精确性和鲁棒性上表现最佳，能够有效预测城市街区的内涝深度及相应的风险等级。研究进一步采用了可解释性算法（SHAP）对LightGBM模型进行解耦分析，结果显示，水文气象因子（降雨总量和降雨延时）及部分建筑配置因子（如建筑密度和建筑拥挤度）为主要的灾害影响因子；另外，水体率对内涝的调节和蓄积起到重要作用，特别是当其超过2.5%时，表现出显著的内涝抑制效果。本研究为城市内涝预测提供了新的技术方法，并从建成环境视角揭示了影响城市内涝的因素以及其内在机制，对高密度城市韧性的提升具有重要的科学意义。

关键词

城市内涝；机器学习；模型性能评估；比较研究；可解释性分析；高密度城市

文章亮点

- 提出结合LightGBM模型和SHAP可解释性算法的综合研究框架，并预测城市街区的内涝深度及相应的风险等级
- 通过机器学习方法证实高密度城市中的老城区在极端暴雨中内涝风险更高
- 首次针对高密度城市，从水文气象、城市下垫面和建筑配置三大类因子出发，探讨城市内涝影响因素和内在机制

基金项目

广东省自然科学基金青年提升项目“基于气候适应性的城市灰绿基础设施韧性增强及动态规划”（编号：2023A1515030158）

编辑 高雨婷，田乐

1 引言

城市韧性建设的核心任务之一是在空间规划阶段精确且有效地预测潜在的城市风险及其影响，并针对性地提出适应性规划策略[1]。随着人工智能技术的不断进步，基于数据驱动的机器学习技术已在城市内涝风险预测方面获得了广泛应用[2]-[5]。埃勒姆·拉菲伊—萨杜伊等人使用支持向量机模型绘制了伊朗希亚夫—柴流域的洪水脆弱性图[3]；王兆礼等人基于随机森林模型评估了东江流域的洪水风险[4]。与传统水文、水力模型相比，基于机器学习的模型的一大优势在于能够在计算资源有限的情况下处理复杂的高维数据，特别是解析多变量与目标值之间的非线性关系[5]。然而，由于过拟合（数据量大时易陷入局部最优解）和计算问题（数据结构复杂时难以找到最佳方案），传统机器学习模型在实际应用中仍存在较大不确定性。

近年来，一种鲁棒性超越了传统机器学习算法的新型集成机器学习模型应运而生，并被广泛应用于城市水文管理等领域[2][6]。侯赛因·沙菲扎德—莫加达姆等人分别使用传统机器学习模型和集成机器学习模型进行洪水敏感性预测，发现集成机器学习模型的预测更准确且表现更稳定[7]；郭宇晨等人证实了集成机器学习模型在洪水预测方面明显优于反向传播深度学习神经网络这一传统机器学习模型[8]；吴泽宁等人应用集成机器学习模型梯度提升树算法预测城市洪水内涝深度，证实了其具有较高的准确率[9]。但当前研究多聚焦于探讨某一机器学习算法在城市洪涝预测中的实用性，对集成机器学习模型在城市多场景内涝预测中的探讨和模型的详细比选关注较少，且较少将其应用到空间层面。本研究旨在将集成算法LightGBM（Light Gradient Boosting Machine）[10]与三种主流传统机器学习算法——随机森林（Random Forests，RF）[11]、支持向量回归（Support Vector Regression，SVR）[12]及反向传播深度学习神经网络算法（Backpropagation Deep Neural Networks，BPDNN）[8]——进行详细对比分析，以揭示这四种算法在预测高密度城市内涝风险方面的性能差异，同时精确解析其影响因素。

从建成环境的角度出发，本文提出了一系列针对城市韧性提升的建设性建议，为高密度城市内涝风险预测提供了创新方法，从理论和实践层面为未来城市规划提供了有益的启示和指导。

2 研究区域与研究方法

本研究选取深圳市作为高密度城市的典型代表，以2019年1月1日至2021年12月31日的历史降雨事件为样本。研究首先构建多因子的联动数据集，涵盖水文、气象、城市形态和内涝事件等多方面数据用于模型的训练与测试。其中，解释性变量包括共计21个自变量的三大类因子（水文气象因子、城市下垫面因子和建筑配置因子），而目标变量为内涝深度数据。研究进一步比较了LightGBM、RF、SVR和BPDNN四种机器学习模型在预测城市内涝风险方面的性能差异。基于对模型精确性和鲁棒性评估，得到最优模型，并绘制深圳市内涝风险易感性分布图。最后，基于预测结果，研究运用可解释性算法（shapley additive explanations，SHAP）从全局特征重要性、特征依赖性及局部特征解释性分析三个维度进行了深入分析，并为提升城市韧性提供了实用的决策参考（图1）。

2.1 研究区域

深圳市地处广东省南部、珠江口东岸，属于典型亚热带季风性气候，长夏短冬，雨量充沛。截至2022年底，全市下辖9个行政区和1个功能区（大鹏新区），总面积为1 997.47km^2[13]，常住人口约1 766万人[14]。

研究表明，深圳市的内涝灾害主要由夏季的瞬时暴雨引发[15]。统计数据显示，全市年降水量平均达1 932.9mm，其中全年约86%的降雨量集中在4~9月[16]。近年来，伴随着城市建成区的扩张，城市空间结构与下垫面状况大幅改变，城市绿地和蓄水设施空间不断被挤占，加剧了城市内涝的风险。

2.2 数据来源与预处理

2.2.1 城市内涝深度数据

研究使用的城市内涝深度数据来自深圳市171个城市内涝监测站（2019~2021年）。数据由路肩附近安装的膜压传感器收集，采样间隔为1小时（图2）。综合考虑地理和水文特性，研究区域被划分为171个次集水单元；以2019年1月1日至2021年12月31日的历史降雨事件（共26 305个样本）作为模型的原始样本数据，按照降雨延时（LTIME）为12小时划分降雨等级，从连续的时间序列中提取出独立的降雨事件[17]并移除总降雨量小于1mm的无效降雨事件。研究筛选出167个可以准确代表对城市内涝有显著影响的有效降雨事件汇编为最终样本数据集。

2.2.2 影响因子数据

现有文献已表明水文气象[18]~[20]、城市下垫面[4][21]~[24]及建筑配置[4][25]~[27]是影响城市内涝生成和发展的主要因素。因此，研究共选择了21个自变量——包括2个水文气象因子[15][28]、10个城市下垫面因子[24][28]~[31]（表1）、9个建筑配置因子[25][29][32]~[34]（表2）——作为预测模型的输入特征，并对各次集水单元分别进行统计。

2.2.3 设计降雨情景

随着全球气候变化的不断加剧，未来城市可能会更频繁地受到极端天气的影响，导致城市内涝风险升高。在评估整个地区的内涝风险时，有必要对不同的降雨情景进行模拟预测。本研究设定降雨持续1小时，重现期为1年、2年、3年、5年、10年和20年，以此将167个有效降雨事件划分为6种降雨情景，带入深圳暴雨强度公式[35]计算暴雨强度：

$$q = \frac{8.701(1 + 0.594)\lg R}{(t + 11.13)^{0.555}}, \qquad (1)$$

式中，q为雨量，t为降雨延时，R为重现期。在其基础上，运用最接近实际观测条件的芝加哥雨型法进行模拟[18][36]，并输出不同设计重现期的降雨量数据。

2.3 研究模型与方法

本研究对4种典型的机器学习算法模型进行超参数优化，通过训练和验证后，应用常用的模型评价指标（R^2、*MAPE*和*RMSE*）对比检验4种模型的表现，并得到预测高密度城市内涝深度的最佳模型；同时，研究使用SHAP从全局特征重要性、特征依赖性及局部特征解释性展开分析，确定影响城市淹没的关键因子，以此提出改善城市内涝问题的针对性建议。

2.3.1 机器学习算法模型

（1）LightGBM

集成算法LightGBM是微软在梯度提升决策树（GBDT）的基础上提出的一种分布式梯度提升算法[37]，也是目前最高效的机器学习算法之一[10]；其在原有的GBDT的算法结构上提出了基于梯度的单边采样算法（GOSS）、互斥特征捆绑算法（EFB）和基于直方图的决策算法三大改

表 1：水文气象因子和城市下垫面因子自变量

因子类型	自变量	数据来源	空间分辨率（m）
水文气象	降雨延时（LTIME）	深圳市气象局	—
	总降雨量（TOTAL_R）		
城市下垫面	绿地率（PGS）	全球 30 米地表覆盖（GlobeLand 30）数据集	30×30
	水体率（PW）		
	不透水面积比（PIS）		
	道路面积率（PR）		
	平均坡度（AS）	地理空间数据云	30×30
	平均海拔高度（AA）		
	平均粗糙度（ARH）		
	平均地形起伏（ARF）		
	与河流的距离（DR）	谷歌地图	—
	归一化植被指数（NDVI）	美国国家航空航天局	250×250

表 2：建筑配置因子自变量

自变量	描述	单位	公式	数据来源
建筑密度（*DB*）	某一范围内每公顷的建筑物数量	n/hm^2	$DB = \frac{N}{A}$	OpenStreetMap
平均建筑高度（*MBH*）	某一范围内建筑的平均高度	m	$MBH = \frac{\sum_{i=1}^{N} H_i}{N}$	
平均建筑体积（*MBV*）	某一范围内建筑的平均体积	m^3	$MBV = \frac{\sum_{i=1}^{N} V_i}{N}$	
建筑高度标准差（*SDBH*）	某一范围内建筑物高度的平均变化	m	$SDBH = \sqrt{\frac{\sum_{i=1}^{N}(H_i - MBH)^2}{N}}$	
建筑体积标准差（*SDBV*）	某一范围内建筑物体积的平均变化	m^3	$SDBV = \sqrt{\frac{\sum_{i=1}^{N}(V_i - MBV)^2}{N}}$	
容积率（*FAR*）	某一范围内所有建筑面积总和与该范围总面积的比值	—	$FAR = \frac{\sum_{i=1}^{N}(F_i \times S_i)}{A}$	
建筑覆盖率（*BCR*）	某一范围内所有建筑首层面积之和与该范围总面积的比值	—	$BCR = \frac{\sum_{i=1}^{N} S_i}{A}$	
建筑形状系数（*BSC*）	某一范围内建筑物与室外大气接触的外表面积与其体积的比值	m^{-1}	$BSC = \frac{\sum_{i=1}^{N} \frac{P_i \times H_i + S_i}{V_i}}{N}$	
建筑拥挤度（*BCD*）	某一范围内所有建筑体积之和与最高建筑体积的比值	—	$BCD = \frac{\sum_{i=1}^{N} V_i}{\max(H_i) \times A}$	

注

N为某一范围内建筑物的数量；A为某一范围的面积；H_i为某一范围内第i栋建筑物的高度；F_i为第i栋建筑物的层数；S_i为第i栋建筑物的建筑面积；P_i为第i栋建筑物的周长；V_i为第i栋建筑物的体积（来源：参考文献 [25][34]）。

进策略[38]。在给定训练数据集$\{(x_i, y_i)\}i=1^N$（x_i为自变量，y_i为目标变量）的基础上，LightGBM算法的目标是最小化特定损失函数的预期值，目标函数定义如下：

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \ , \qquad (2)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \ , \qquad (3)$$

式中，Obj是要最小化的目标，$l(y_i, \hat{y}_i)$是每个样本的目标值y_i和预测值$\hat{y}_i$的损失函数。为了防止过度拟合，以公式（3）定义正则化项，其中，$\Omega(f_k)$是正则化项，f_k是第k个树模型，γ是加入新叶子节点引入的复杂度代价，T是树的叶子节点数，λ是叶子权重修正系数，ω是叶子权重值。

（2）RF

RF是最初由里奥·布雷曼于2001年提出[11]的集成学习模型，它是袋式决策树的改进版，最终的分类或回归是通过对所有单个决策树使用多数投票方案来完成的[39]，是一种灵活且易于使用的机器学习算法。

（3）SVR

作为SVM的一个分支，SVR主要用来解决回归问题。SVR的目标是要使得超平面与最远样本点的距离最小，从而可以利用超平面对数据进行拟合，而这一过程需要选择合适的核函数。一些研究表明，径向基函数（RBF）在城市内涝预测方面优于其他核函数[12][40]。

（4）BPDNN

BPDNN是一种根据输出和期望输出之间的误差反向传播算法训练的多层前馈网络，因其具有多层结构的特征而可用于处理非线性问题，BPDNN已被广泛运用在洪涝灾害风险评估中。[8]

2.3.2 模型性能评估方法

本研究采用R^2、*MAPE*和*RMSE*作为评价指标来评价和比较模型的性能。R^2衡量回归模型的拟合程度[41]，*MAPE*衡量预测值与真实值的相对误差大小[42]，*RMSE*评估模型预测值与真实值之间的平均误差[43]，三者的数学公式如下：

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad , \tag{4}$$

$$MAPE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| , \tag{5}$$

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad , \tag{6}$$

式中，y_i指第i个被观测到的内涝深度数据，$\hat{y}_i$指相应的预测城市内涝深度数据（cm），n代表观测次数。

2.3.3 SHAP算法

难以解释（即黑盒问题）一直是集成机器学习的弊端之一[44]。SHAP算法基于合作博弈论构建了一个加性的解释模型，提供了一种估计每个特征贡献度的方法。近些年已经有不少研究开始使用SHAP对复杂的集成模型进行可视化解释，结果证明其具有较强的可信度[45][46]。本研究采用SHAP算法探究各特征与城市内涝深度的全局关系及其重要性，同时进行局部特征解释性分析。其解释模型如下：

$$g(z') = \phi_0 + \sum_{i=1}^{M}\phi_i z_i' \quad , \tag{7}$$

式中，g是解释模型，$z' \in \{0, 1\}^M$是联盟向量，M是最大的联盟规模，ϕ_i是特征i对模型输出的贡献。

在SHAP算法中，每个特征的贡献度是根据他们的边际贡献来分配的[47]，主要通过计算特征对模型输出的边际贡献，再从全局和局部两个层面对机器学习模型进行解释；Shapley值不仅可以反映每个特征的重要性，还可以反映其对目标值的正负影响[48]。

3 结果与讨论

3.1 算法模型性能比较

对训练集模型性能的评估结果表明（表3），LightGBM（R^2=0.96）的拟合程度明显优于RF（R^2=0.89）、SVR（R^2=0.78）和BPDNN（R^2=0.69）；同时，LightGBM展现出最高的拟合精度（*MAPE*=0.22），相对误差最小，其次是RF（*MAPE*=0.32）、SVR（*MAPE*=0.44）和BPDNN（*MAPE*=0.63）；在*RMSE*评价指标上，LightGBM也表现出最低值（*RMSE*=1.81cm），偏差最小，预测精度表现最好，随后依次是RF（*RMSE*=2.26cm）、SVR（*RMSE*=3.12cm）和BPDNN（*RMSE*=3.64cm）。总体而言，LightGBM的精确性和稳定性最高。

为验证模型在预测城市内涝深度方面的鲁棒性，依据机器学习算法逻辑选取了25%（42场）的降雨事件数据（共计1 790个样本），作为独立测试数据。结果表明，LightGBM在测试集中依旧是预测城市内涝深度性能最佳的模型（R^2=0.90、*MAPE*=0.28、*RMSE*=2.29cm）。

结果显示，LightGBM和RF结果的拟合线更接近“y=x”的理想拟合线（图3）。进一步对比分析发现，虽然RF在预测内涝深度小于10cm时表现与LightGBM基本相当，但当深度大于10cm时，内涝深度值越大，LightGBM的优越性越显著。综合指标评价结果及绝对误差点分布图，可以得出结论：在预测精度和鲁棒性上，LightGBM明显优于其他三种机器学习模型。因此，本研究利用LightGBM预测深圳市不同降雨情况下的城市内涝深度，并绘制相应的内涝风险地图。

3.2 城市内涝风险预测结果

本研究根据自然断点法将次集水单元的内涝深度峰值风险划分成5种风险区，即极低风险区（0~5cm）、低风险区（5~10cm）、中风险区（10~15cm）、高风险区（15~20cm），以及极高风险区（≥20cm）。同时设定降雨延时为1小时，比较不同重现期（1年、2年、3年、5年、10年和20年）的城市内涝风险变化情况（图4）。

表 3：四个模型在训练集和测试集中的性能评估

模型	R^2		*MAPE*		*RMSE* (cm)	
	训练集	测试集	训练集	测试集	训练集	测试集
LightGBM	0.96	0.90	0.22	0.28	1.81	2.29
RF	0.89	0.81	0.32	0.43	2.26	3.15
SVR	0.78	0.43	0.44	0.51	3.12	4.25
BPDNN	0.69	0.67	0.63	0.77	3.64	4.07

注

R^2 越接近 1 表示模型拟合程度越高；*MAPE* 越接近 0 表示模型预测模型越准确；*RMSE* 越接近 0 表示模型预测误差越小，模型的预测能力越强。

随着降雨重现期由1年增至10年，城市内涝极低风险区的面积占比逐渐减小，而低风险区和中风险区的面积占比相应增加（图4）。而在此阶段，未观察到高风险区和极高风险区的形成。结果表明，尽管在某些排水设施落后的区域可能出现局部积水，深圳市大部分区域的排水系统可以有效应对中等降雨事件。然而，当重现期增至20年时，城市内涝风险等级的空间分布发生了明显变化，特别是中风险区的面积占比显著增加。这表明在遭遇20年一遇的降雨事件时，深圳市大部分次集水单元或将会面临不同程度的城市内涝，其中高风险区主要分布在排水系统和地形复杂的老城区（南山区、福田区及坪山区）。

3.3 模型可解释性分析

3.3.1 全局特征重要性

全局特征重要性是一种衡量模型中每个特征对预测结果影响大小的指标，主要通过每个变量Shapley值的绝对平均值进行评估[49]。通过对每个特征进行计算（图5），研究发现在排序前60%的特征中，水文气象因子TOTAL_R和LTIME对模型的影响极为显著，重要性贡献度分别为24.6%和15.8%。其次，城市下垫面因子PW及PIS的重要性贡献度分别为6.4%和5.1%。而建筑配置因子中，*BCD*和*DB*的重要性贡献度分别为4.4%和4.1%。

整体来看，水文气象因子对城市内涝深度的影响显著超过了其他两类因子，这也验证了之前多项研究将降雨量和降雨持续时间等水文气象因子视为城市内涝主要致灾因子的结论[18]-[20]。另一方面，PW是上述6个指标中唯一对城市内涝有抑制作用的因子，这在一定程度上解释了许多城市选择在城区内修建人工湖的原因——人工湖能够在极端降雨事件中蓄集地表径流，进而减轻下游受纳水体和雨水管网压力。

在建筑配置因子中，*BCD*和*DB*均描述了建筑密度，城市中的PIS随着二者的增加而增加并导致可蓄洪的蓝绿基础设施相应减少，进而增加城市内涝发生的概率。

3.3.2 特征依赖性

为深入探索城市内涝深度与主要致灾因子之间的非线性关系，本研究采用SHAP算法的特征依赖性图来揭示单个变量对模型输出的贡献——特征依赖性图可展示某特征对预测模型输出的改变程度，以清晰的可视化方式阐释特征值与相应的Shapley值之间的边际效应。Shapley值的正负标志代表特定特征对预测结果的积极或消极影响[50]。本研究依据全局重要性分析的结果，从三大特征类别中选取排名最靠前的6个因子进行分析。

（1）水文气象因子

结果显示，TOTAL_R与城市内涝深度大体呈正相关关系（图6-1）。当TOTAL_R低于25mm，其对城市内涝的影响相对较弱。然而，当TOTAL_R超过25mm，其对城市内涝的影响可体现为三个阶段。首先，在降雨总量从25mm升至100mm的阶段，城市内涝风险迅速增加，Shapley均值从2提升至6左右；在100mm至125mm的阶段，降雨的影响保持稳定，降雨量的增加未引发更严重的城市内涝，此时呈现边际效应；当TOTAL_R超过125mm时，Shapley值均值再度快速上升，城市内涝风险继续增加，这可能与已经接近或超过城市排水系统处理能力的极限有关。尽管各地的城市环境和基础设施配置差异可能导致各地对降雨量的敏感性有所不同，但降雨量对内涝风险强有力的影响在各地都是一致的。

LTIME的Shapley值依赖图显示（图6-2），当持续降雨小于55小时，LTIME对模型输出的产生了波动影响。尽管降雨持续时间的延长会加剧城市内涝，但并未出现类似TOTAL_R的线性增长影响。然而，当降雨延时超过55小时，Shapley值急剧增加。研究认为，当LTIME小于55小时，大部分降雨事件的每小时降雨量并没有达到致灾阈值，因此对城市内涝深度的提升并不强烈。

（2）城市下垫面因子

城市下垫面因子的特征依赖性图显示（图7-1），PW能有效降低城市内涝风险：水体面积占比越大，对城市内涝的削弱作用越显著；在达到一定阈值（12.5%）之后，内涝深度受其影响不明显。当PW小于1.2%时，水体蓄集洪水的效果不明显，但当水体率大于2.5%时，其对城市内涝的削弱效果逐渐增强。齐文超等人提出，城市中的湖泊在汛期可为地表径流排放提供大面积的缓冲区，从而提高城市内涝防治能力[51]。因此，保护自然湖泊和合理开挖城市人工湖对城市雨洪管理有积极影响。

PIS对城市内涝的影响主要分为三个阶段（图7-2）：1）从0到15%，在此阶段内Shapley值始终保持在0附近，对城市内涝影响较小；2）当PIS超过15%的阈值时，会逐步加剧城市内涝风险，并在30%时达到最高水平；3）当PIS高于30%时，并未继续加剧城市内涝风险①。相关研究表明，不透水表面的空间配置对城市内涝具有重要影响。降低PIS可为土壤吸纳雨水或植被拦截雨水提供足够的空间，有助于降低道路上的峰值径流，从而抑制城市内涝[52]。

（3）建筑配置因子

在建筑配置方面，*BCD*对城市内涝深度有压迫影响（图8-1），且观察到当*BCD*在[0.01，0.02]的区间内，Shapley值均值突然增加。对比样本数据可知，这主要是由于在这些次集水单元中存在大面积硬质铺装的空地，尽管其*DB*不高，但大面积的不透水表面阻碍了降雨的有效下渗和蓄积，造成短时间内地表径流增大，从而引发城市内涝。基于此，在城市设计或城市更新的过程中，需要严格控制不透水铺装的占比。当*BCD*超过0.08的阈值时，会对内涝产生更为明显的压迫作用。而*DB*对城市内涝

① 当 PIS 高于 30%，其对内涝的影响并未加剧，故图 7 中重点展示 0~30% 的区间。

的影响具有明显空间异质性（图8-2），相同的*DB*条件下，建筑布局、建筑高度及其形态等其他空间配置的差异会导致Shapley值存在较大的不同：当*DB*小于15/hm^{2}时，Shapley值的波动区间是[－0.5，1.0]；而当*DB*值大于15/hm^{2}时，波动区间变为[0.0，2.0]，这意味着一旦超过15/hm^{2}这一阈值，*DB*的提高将会增加城市内涝风险。

3.3.3 局部特征解释性分析

通过分析每个次集水单元中不同特征对模型输出的贡献度，可深入理解空间异质性对城市内涝的影响，为后续的城市内涝适应性策略提供针对性建议。根据先前的预测结果（图4），研究选择了两个典型的深圳市建成环境城市内涝高风险的次集水单元（A代表高密度老城区，B代表沿海新开发地区），并进行单一样本的特征研究（图9）[②]。

对于单元A，占贡献度前80%的特征分别为TOTAL_R、*BCR*、PIS和*MBV*。与上文全局特征重要性分析结果一致，TOTAL_R仍然是影响该地区城市内涝风险的最重要因素。单元A作为深圳老城区的典型代表，*DB*极高，道路系统复杂，大部分场地为不透水表面，易于在极端暴雨中形成大量地表径流。

在单元B中，占贡献度前80%的特征为TOTAL_R、*MBV*、*BCR*、PIS和LTIME；此外，或因两个地块PGS明显不同，相较于单元A，PGS的贡献度明显提升。这可能是由于单元B位于南山区沿海的商业区和高档住宅区，尽管建筑分布依然密集，但场地中沿海岸线设有一些带状绿地、湿地公园和街边绿地，因此PGS在模型贡献度上的作用得到凸显，对城市内涝起到了抑制作用。

4 结论和展望

本研究以深圳市2019～2021年的城市内涝事件观测数据为依据，综合水文气象、城市下垫面，以及建筑配置三大类指标，运用机器学习模型预测城市内涝风险。通过比较LightGBM、RF、SVR以及BPDNN四种模型的模型性能，最终选定LightGBM作为最佳预测模型进行城市内涝风险评估。

研究结果表明，当深圳市遭遇高重现期（如20年一遇）的降雨，将出现城市内涝高风险区，主要集中于南山区、福田区及坪山区的老城区。水文气象因子（包括TOTAL_R和LTIME）是主要的灾害影响要素，对模型贡献度达到了40.4%。当TOTAL_R超过125mm或LTIME超过55小时，对城市内涝的影响尤为显著。PW是唯一对城市内涝有抑制作用的指标，当其面积占比超过2.5%时，对洪涝的调节和滞蓄作用逐渐增强。与此同时，建筑密度相关特征在超出特定阈值后，与城市内涝风险表现出显著正相关关系。值得注意的是，ARF等下垫面因子的解释性均较低，这可能与深圳市建成区地形起伏落差较小，且不透水率较高有关。此外，TOTAL_R和*DB*仍是影响局部地区内涝深度的主要因素，而在特定区域，蓝绿基础设施对于城市内涝抑制的作用十分关键。

本研究提出并论证的基于LightGBM算法的城市内涝风险预测方法具有普适意义；同时，可通过可解释性算法分析，获取影响城市内涝的关键因子，以为城市规划与建设提供指导。在高密度城市建设发展过程中，建议加大对老城区的更新改造力度，修复自然生态系统，抓紧补齐排水防涝设施短板，提高排水和防洪能力；适当恢复并增加水体空间，扩展城市及其周边的自然调蓄空间，按照有关标准和规划开展滞洪蓄洪空间和安全工程建设；同时，在城市建设和更新中结合空间和竖向设计留白增绿；提高硬化地面中可渗透面积比例，因地制宜使用透水性铺装；积极推进海绵城市建设，保护城市山体，修复江河、湖泊、湿地等，保留天然雨洪通道、滞洪蓄洪空间，构建连续完整的生态基础设施体系。

受城市管网数据获取的限制，本文采用了相关道路因子（如PR）作为代理变量[28]，虽然其可在一定程度上反映城市排水的效率，但也存在局限性。使用空间分辨率更高的数据将有助于更细致地揭示城市内涝动态，进而更准确地识别和分析影响城市内涝的关键因素。未来还可以结合水文水力模型，提升实验的针对性和精确性。

② 因排名靠后的特征贡献值过小且不具有进一步讨论的典型性，故局部特征解释性分析聚焦贡献度前 7 位的特征。

图 1. 模型比较及城市内涝风险预测研究框架
图 2. 深圳市 171 个城市内涝监测站分布图（数据来源：深圳市气象局）
图 3. 基于测试数据集的预测及实际城市内涝深度的散点和拟合线图（红线为拟合回归线，红色区域代表 95% 的置信区间）
图 4. 不同重现期下降雨延时 1 小时的城市内涝风险易感性分布图
图 5. LightGBM 模型特征 Shapley 值的特征散点图（左图）和全局特征重要性排名（右图）
图 6. 重要变量和 Shapley 值的依赖关系图（黑线为区分不同阶段的辅助线）：TOTAL_R 的 Shapley 值依赖图（图 6-1）；LTIME 的 Shapley 值依赖图（图 6-2）。
图 7. 重要变量和 Shapley 值的依赖关系图：PW 的 Shapley 值依赖图（图 7-1）；PIS 的 Shapley 值依赖图（图 7-2）。
图 8. 重要变量和 Shapley 值的依赖关系图：*BCD* 的 Shapley 值依赖图（图 8-1）；*DB* 的 Shapley 值依赖图（图 8-2）。
图 9. 单元 A 和单元 B 位置图（左图）和局部特征贡献排名图（右图）