

Article

Explainable Machine Learning-Based Urban Waterlogging Prediction Framework

Yinghua Deng and Xin Lu * 

College of Systems Engineering, National University of Defense Technology, Changsha 410073, China; yinghua.deng@outlook.com

* Correspondence: xin.lu.lab@outlook.com

Abstract

Urban waterlogging has become a critical challenge to urban sustainability under the combined pressures of rapid urbanization and increasingly frequent extreme weather events. However, traditional predictive models struggle to achieve real-time, point-specific early warning effectively, primarily due to the interference of redundant high-dimensional data and the inability to handle severe data imbalance. This study proposes a lightweight and interpretable machine learning framework for real-time waterlogging hotspot prediction, based on a multi-dimensional feature space. Specifically, we implement a Lasso-based mechanism to distill 37 multi-source variables into five core determinants. This process effectively isolates dominant environmental drivers while filtering noise. To further overcome the recall bottleneck, we propose a Synthetic Minority Over-sampling Technique based on Weighted Distance and Cleaning (SMOTE-WDC) algorithm that incorporates weighted feature distances and density-based noise cleaning. Validating the framework on datasets from Shenzhen (2023–2024), we demonstrate that the integrated Gradient Boosting Decision Tree (GBDT) model integrated with this strategy achieves optimal performance using only five features, yielding an F1-score of 0.808 and an Area Under the Precision-Recall Curve (AUC-PR) of 0.895. Notably, a Recall of 0.882 is attained, representing a 4.6% improvement over the baseline. This study contributes a cost-effective, high-sensitivity approach to disaster risk reduction, advancing predictive urban waterlogging management.

Keywords: urban waterlogging; prediction; explainable machine learning; crisis management

1. Introduction

Amid ongoing global warming and rapid urbanization, urban waterlogging has emerged as a growing threat, posing significant challenges to urban safety and socio-economic development [1–3]. Pluvial flooding occurs when rainfall intensity exceeds the design capacity of urban drainage systems, resulting in surface water accumulation that cannot be discharged promptly. This phenomenon not only causes traffic disruptions, infrastructure damage, and environmental degradation but also endangers residents' lives and property, severely impacting population mobility, urban functionality and sustainable development [4–9]. Enhancing the predictive capabilities and emergency response mechanisms for urban waterlogging has therefore become an urgent priority in urban governance.

Urban waterlogging disasters arise from the combined effects of natural factors and human activities [10,11]. To achieve effective flood prediction and management, it is essential to identify key influencing factors. In terms of natural influencing factors, meteorological conditions play a particularly prominent role. Sudden surges in rainfall intensity or the



Academic Editor: Milan Gocić

Received: 19 January 2026

Revised: 2 March 2026

Accepted: 6 March 2026

Published: 13 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

occurrence of extreme storm events can easily exceed the load-bearing threshold of urban drainage systems, thereby inducing urban waterlogging disaster [12,13]. Beyond meteorological drivers, the physical properties of urban land, especially topographic elevation and slope, play a pivotal role in the occurrence of pluvial flooding [14]. Water flows tend to converge toward low-lying zones by gravity, causing a sharp surge in water discharge within a short period and thus substantially elevating flood risks in downstream lowlands [15,16]. Regions such as underground parking lots and basements are particularly susceptible to water accumulation and severe impacts [17].

Notably, urban waterlogging is not exclusive to low-lying regions, during intense rainfall events, higher-elevation areas may also experience inundation, primarily due to inadequate drainage infrastructure (e.g., clogged pipes or insufficient pumping capacity) that fails to discharge excess runoff in a timely manner. As urbanization accelerates, the increase in impervious surfaces (e.g., roads and buildings) significantly reduces rainfall infiltration, enhancing surface runoff and raising the risk of urban waterlogging [18–21], whereas the presence of green spaces and water bodies helps mitigate this risk [22,23]. Furthermore, vertical urban features—such as building characteristics and spatial layouts—can alter runoff pathways and flow velocities, which in turn exerts a significant influence on the spatial distribution of pluvial flooding [24–26]. The interplay between these topographical, hydrological, and anthropogenic factors is typically non-linear and highly complex, creating challenges for the accurate prediction of urban waterlogging events and the formulation of effective response strategies [27,28].

To empirically illustrate how the aforementioned natural and anthropogenic factors synergistically exacerbate urban waterlogging, a series of high-impact global disasters serve as compelling examples. For instance, during Hurricane Harvey in Houston (2017), the city's low-lying topography combined with extensive impervious surfaces from urban sprawl severely hindered rainwater runoff and infiltration. This led to waterlogging depths exceeding 1 m in vast areas, inflicting direct economic losses of \$125 billion and ranking it among the costliest disasters in U.S. history [29]. Similarly, the record-breaking rainfall event in Rio de Janeiro (April 2019)—with 24 h accumulations peaking at 343.4 mm—overwhelmed the city's drainage systems, and the uneven elevation of built-up areas directed concentrated runoff toward densely populated neighborhoods, triggering widespread waterlogging [30]. Most notably, the “7·20” extreme rainfall in Zhengzhou (2021) exposed the critical limitation of inadequate drainage infrastructure in coping with extreme precipitation, which, coupled with the city's sprawling impervious road and building networks, resulted in catastrophic urban waterlogging, 120.06 billion RMB in direct losses, and 398 deaths [31]. These cases demonstrate that flood-induced urban waterlogging is becoming an increasingly severe global threat [32]. Therefore, enhancing accurate prediction and real-time early warning of urban waterlogging is imperative. Particularly for urban decision-makers, improving urban waterlogging management capabilities enables authorities to take emergency actions to prevent casualties and reduce damage.

In recent years, a variety of methods have been developed to simulate and predict urban waterlogging, which can be broadly categorized into physics-based models and data-driven models. Physics-based models, such as SWMM (Storm Water Management Model), Mike Urban, and InfoWorks ICM (Integrated Catchment Modeling), simulate urban waterlogging by modeling physical processes such as rainfall, surface runoff, and drainage networks [33], and are widely applied in optimizing urban drainage systems [34] and assessing urban waterlogging risks [35]. The SWMM model, a classic tool for urban drainage system design and evaluation, simulates the generation, transmission, and operation of rainfall runoff in detail. It calculates the formation and flow of surface runoff accurately by considering factors such as rainfall intensity, surface type, and slope [36]. The Mike Urban

model, on the other hand, focuses on hydraulic and hydrological simulations of urban drainage systems, effectively modeling complex drainage networks and surface water flow processes. By coupling 1D and 2D flow simulations, Mike Urban offers a comprehensive view of urban flood dynamics and is extensively used in flood risk assessments to identify flood-prone areas and key drainage bottlenecks. The InfoWorks ICM model integrates watershed hydrological processes with urban drainage system simulation, providing a comprehensive hydrological analysis from the watershed to the urban scale, strengthening the scientific and targeted urban water resource management [37]. These physics-based models have significant advantages in urban flood simulations, as they can describe the physical processes of rainfall runoff in detail, providing high-precision simulation results suitable for flood modeling at different scales. However, these models are often computationally intensive, making them challenging to meet the real-time prediction needs of flood disaster management. Furthermore, they require high-precision data on topography and drainage networks, which are challenging to acquire. Moreover, the uncertainty of model parameters can affect prediction results, thereby imposing certain limitations on their prediction accuracy and applicability.

Compared to traditional physical models, data-driven models demonstrate substantial advantages in computational efficiency and predictive accuracy, constituting a data-driven and information-oriented paradigm [38] that effectively processes complex multi-source datasets with a reduced computational burden, thereby enabling more timely and precise short-term urban waterlogging forecasting [39,40]. By leveraging machine learning techniques, these models analyze vast repositories of historical data to capture the complex, non-linear relationships between driving variables and urban waterlogging processes at a lower computational cost. Researchers have explored various machine learning algorithms in the field of urban waterlogging prediction, such as Artificial Neural Networks (ANN) [41,42], neuro-fuzzy [43,44], Support Vector Machines (SVM) [45,46], and Gradient Boosting Decision Trees (GBDT) [47], all of which have been proven effective for both short-term and long-term forecasting. By learning intricate patterns from historical data, these methods capture the dynamic relationships between rainfall, runoff, and drainage systems, thereby improving the accuracy and timeliness of urban waterlogging predictions. Despite the significant advancements brought by artificial intelligence-based urban waterlogging prediction technologies, several limitations persist. First, model performance is heavily contingent upon data quality; thus, ensuring the accuracy and integrity of data is paramount. Second, constructing robust predictive models requires extensive sample data and meticulous feature selection, necessitating in-depth analysis of specific urban characteristics. Furthermore, machine learning-based models face challenges in hyperparameter selection and tuning, highlighting an urgent need for algorithmic optimization to enhance predictive capability and adaptability.

While these methods have made significant strides in urban waterlogging prediction, current research focuses primarily on the simulation of water level fluctuations or inundation processes. However, the spatial prediction of discrete urban waterlogging points remains relatively under-explored due to the extreme geomorphological complexity and anthropogenic heterogeneity of urban systems. Such complex physical environments pose inherent challenges for modeling frameworks to accurately perceive and interpret localized features [48]. Accurate identification of these risk points necessitates a comprehensive integration of multi-dimensional drivers—including micro-topography [49–51], drainage infrastructure [51–54], and localized rainfall distribution [55–57]—which exhibit intricate non-linear interdependencies. This necessitates a predictive framework capable of capturing these complex interactions at a fine spatial scale to support targeted decision-making.

To bridge these research gaps, this paper proposes an ensemble learning framework specifically designed for real-time waterlogging hotspot prediction. We construct a multi-dimensional feature dataset integrating meteorology, topography, land use, urban morphology, and demographic distribution to capture the complex heterogeneity of the urban surface. However, to circumvent the overfitting risks associated with high-dimensional feature spaces, a Lasso-regularized selection mechanism is integrated into the framework. This approach distills the multi-source dataset into five pivotal determinants, balancing predictive accuracy with the computational efficiency required for rapid field deployment. This reinforces the premise that practical model efficacy is driven by the informational quality of key predictors rather than the redundant complexity of the feature space, offering a scalable and deployable solution for urban flood management.

Compared with traditional physical modeling, our framework frames the prediction as a binary classification task to capture non-linear interactions efficiently. Crucially, effective early warning requires a strategic balance between safety and efficiency. High sensitivity is essential to minimize the loss of life and property, while reasonable precision is necessary to prevent resource wastage. To achieve this, we synergize the novel SMOTE-WDC sampling algorithm with a recall-prioritized threshold strategy. This approach not only mitigates severe class imbalance but also ensures an optimal trade-off between Precision and Recall, providing a cost-effective and actionable solution for urban emergency management.

2. Data and Methods

This study develops a machine learning-based framework for identifying urban waterlogging hotspots, mathematically formulated as a binary classification task. To comprehensively capture the environmental drivers of inundation, we constructed a candidate feature space spanning three key domains: meteorological observations, land surface attributes, and urban morphological parameters. The model is developed and validated using multi-source data collected from waterlogging events in Shenzhen, China, between 2023 and 2024.

2.1. Dataset

2.1.1. Study Area and Flooding Data Pre-Processing

Shenzhen, located in the southern part of Guangdong Province, China, is one of the country's major economic centers. The city boasts a unique geographical setting, characterized by a coastline extending 260 km, and has a subtropical marine climate with an average annual temperature of approximately 22 °C. The region experiences abundant rainfall, primarily concentrated during the wet season from May to September. These climatic characteristics pose significant challenges related to heavy rainfall during the annual flood season, making urban waterlogging a particularly pronounced issue in Shenzhen. In recent years, the rapid pace of urbanization has led to the extensive construction of impermeable surfaces (such as roads, highways, underground passages, and parking lots), which significantly increases surface runoff and further escalates the risk of urban waterlogging [58].

This study focuses on eight major urban waterlogging events that occurred in Shenzhen from 2023 to 2024, specifically on the following dates: 23 May 2023; 14 June 2023; 29 July 2023; 7 September 2023; 6 April 2024; 20 April 2024; 1 May 2024; and 26 July 2024 (Figure 1). These events resulted in substantial rainfall and widespread water accumulation, severely disrupting the normal operations of the city. To comprehensively capture the spatio-temporal dynamics of these urban waterlogging events, real-time water depth data from 206 flood-prone locations were collected through the Shenzhen Water Authority. These measurement stations are distributed across various regions of Shenzhen, encompassing a range of land use types and topographical conditions, thus providing a holistic

representation of water accumulation across the city. High-precision sensors recorded the water depth at each station with minute-level temporal resolution, covering the period from the onset of rainfall to the complete drainage of water.

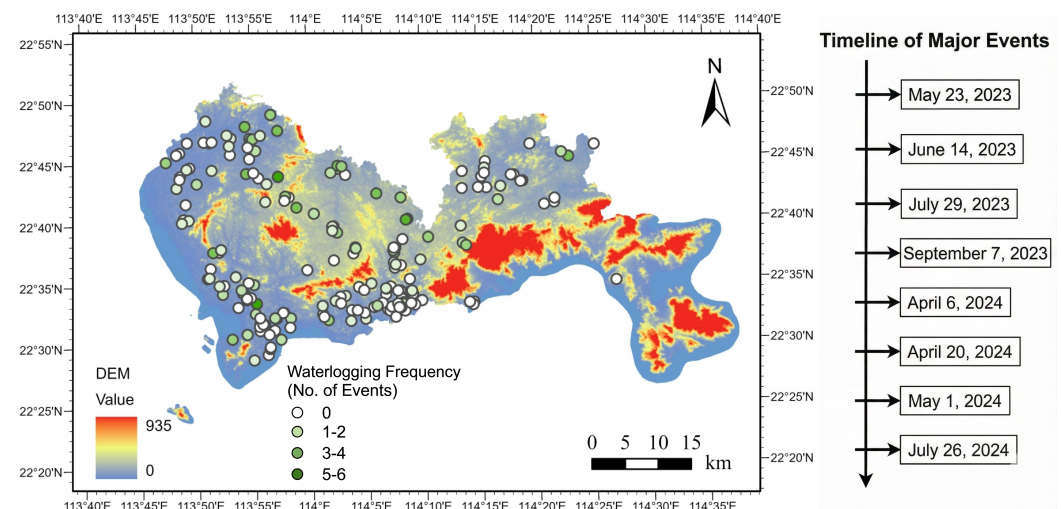


Figure 1. Timeline of waterlogging events in Shenzhen (2023–2024).

To standardize the spatial analysis and facilitate the integration of multi-source data (e.g., meteorological, topographic, and land use data) with water depth observations, the entire study area was partitioned into a 47×93 grid matrix at a 1 km resolution (totaling 4371 units) to ensure complete coverage of the Shenzhen administrative region. This grid-based spatial discretization ensures consistent spatial units for subsequent feature extraction, hotspot identification, and machine learning model training. To support binary classification modeling (waterlogged vs. non-waterlogged), the water depth data were thresholded: grid cells with maximum water depth ≥ 0.15 m during an urban waterlogging event were labeled as 1 (waterlogging), while those with maximum water depth < 0.15 m were labeled as 0 (non-waterlogging). This dataset offers detailed time series information for analyzing the dynamic processes of urban waterlogging and reveals differential responses of various regions to intense rainfall, thereby providing robust data support for subsequent modeling and analysis. The field descriptions of the waterlogging dataset are presented in Table 1.

To ensure the objectivity of model evaluation, a chronological data splitting strategy was employed to maintain temporal independence between the training and testing phases. The first six waterlogging events (from 23 May 2023 to 20 April 2024) were utilized for model construction and cross-validation. The final two events in the dataset (1 May and 26 July 2024) were reserved as an independent blind test set to verify the model's predictive performance on future storm events.

2.1.2. Collection of Land and Built Environment Data

Geographical and built environment characteristics play a critical role in the occurrence and development of urban waterlogging during heavy rainfall events. This study focuses on key factors such as land use, the built environment, and population distribution, which are important in explaining the variations in urban waterlogging severity across different urban areas.

To comprehensively capture these characteristics, relevant data were collected from multiple official sources. Elevation data, sourced from National Geospatial Data with a resolution of 30 m, provides information on terrain variation, slope, and roughness in urban areas. Land use characteristics, including the proportions of green space, water bodies,

impermeable surfaces, and bare land, were obtained from 30-m resolution data provided by Tsinghua University.

In terms of the built environment, this study utilized three-dimensional building metrics, including Mean Building Height (MBH), Mean Building Volume (MBV), and Building Congestion Degree (BCD). These data were sourced from the National Geographic System Science Data Sharing Platform and reflect the spatial distribution and density of urban buildings. Building density and high-rise structures can exacerbate water flow blockage, increasing the risk of urban waterlogging; understanding these features is crucial for identifying potential areas of flooding accumulation.

Drainage capacity is closely related to the density of drainage networks. However, obtaining specific information on urban drainage infrastructure is challenging due to confidentiality. Research indicates that approximately 50% of the street network length corresponds to 80% to 85% of water supply and sewage networks [59]. Therefore, road network density can serve as an alternative indicator of drainage capacity, effectively reflecting the convergence of water flow in urban areas. Particularly during heavy rainfall events, high-density road networks often lead to water retention, increasing urban waterlogging [60,61]. Population density data, sourced from WorldPop with a resolution of 1000 m, are used to capture levels of urban development and population distribution. Higher levels of urban development and population density typically imply greater proportions of impermeable surfaces, potentially increasing urban waterlogging risk.

Table 1. Overview of the multi-source datasets used in this study.

Data	Detail	Secondary Data	Source
Waterlogging hotspots	-	Real-time water depth monitoring data	Water Resources Bureau of Shenzhen
Precipitation	1 km resolution	Maximum Rainfall, Average Rainfall, Maximum Wind Speed, Average Wind Speed, Maximum Humidity, Average Humidity, etc.	Water Resources Bureau of Shenzhen
Building information	10 m resolution	Density of Buildings (DB), Mean Building Height (MBH), Mean Building Volume (MBV), Building Congestion Degree (BCD)	National Earth System Science Data Center
Digital Elevation Model (DEM)	30 m resolution	Elevation, surface relief, roughness, slope	Geospatial Data Cloud
Land use	30 m resolution	Green space, water bodies, impervious surface, bareland	Tsinghua University
Population	1 km resolution	Population	WorldPop [62]

2.1.3. Collection of Meteorological Characteristics

Meteorological conditions are one of the key driving factors for the occurrence of urban waterlogging. During heavy rainfall events, variations in meteorological characteristics such as rainfall amount, rainfall intensity, temperature, atmospheric pressure, wind speed, and humidity directly influence the occurrence and development of urban waterlogging. To comprehensively capture the impact of these meteorological conditions on urban waterlogging events, this study utilized obtained data from the *Shenzhen Automatic Weather Station Real-time Grid Dataset* via the Shenzhen Water Resources Bureau. The utilization of this grid-based dataset is advantageous as it provides uniform spatial coverage and high-resolution granularity (1 km × 1 km), enabling the framework to capture localized, high-intensity rainfall pulses that traditional point-based station data might miss. This dataset is updated daily with a 10-min temporal resolution. The raw dataset contains a variety of meteorological features, including temperature, wind dynamics, air pressure, relative humidity, visibility, and cumulative rainfall at multiple lead times (1 h, 2 h, 3 h, 6 h, and 24 h), among others. Based on these official forecasting records, we further derived

secondary features such as the Rainfall Change Rate to capture the temporal intensity of storms, providing a robust predictive foundation for the prediction model.

The specific meteorological features extracted include maximum and average rainfall over the past 1 h, 2 h, 3 h, and 6 h. Specifically, Max 1 h Rainfall characterizes the immediate impact of short-term flash floods, while Avg 6 h Rainfall serves as a proxy for soil and drainage saturation caused by persistent precipitation. Beyond these standard intensity metrics, this study introduced several derivative features to characterize the temporal instability of weather patterns. Specifically, Rainfall Change Rate was calculated as the ratio of 1 h to 3 h precipitation, with a small epsilon (10^{-6}) added to the denominator to prevent division by zero, effectively identifying sudden cloudbursts. Unlike traditional differencing or smoothing techniques, the raw ratio was preserved to maximize the model's sensitivity to sudden cloudbursts. These rainfall indicators reflect the intensity and duration of precipitation, which directly affects the convergence of water flow and the extent of water accumulation. To account for the cumulative effects and atmospheric trends, we calculated the Mean Lag-1h Sea Level Press and 3 h Moving Avg. Temp., which serve as proxies for atmospheric instability and thermal accumulation. Additionally, Surface Relative Humidity was included to assess the moisture saturation of the near-surface atmosphere. Furthermore, wind dynamics, including Wind Speed, were aggregated to analyze the moisture transport conditions during flood events. To address the non-scalar nature of wind direction, the original degree-based records were transformed using circular encoding (sine and cosine components) to maintain the trigonometric continuity of the vector space. By integrating these multi-source data, this study constructed a multidimensional dataset, which serves as a rich information foundation for predicting urban waterlogging locations using machine learning and also provides comprehensive support for model training and validation. Detailed information on these data is shown in Table 1 and illustrated in Figure 2.

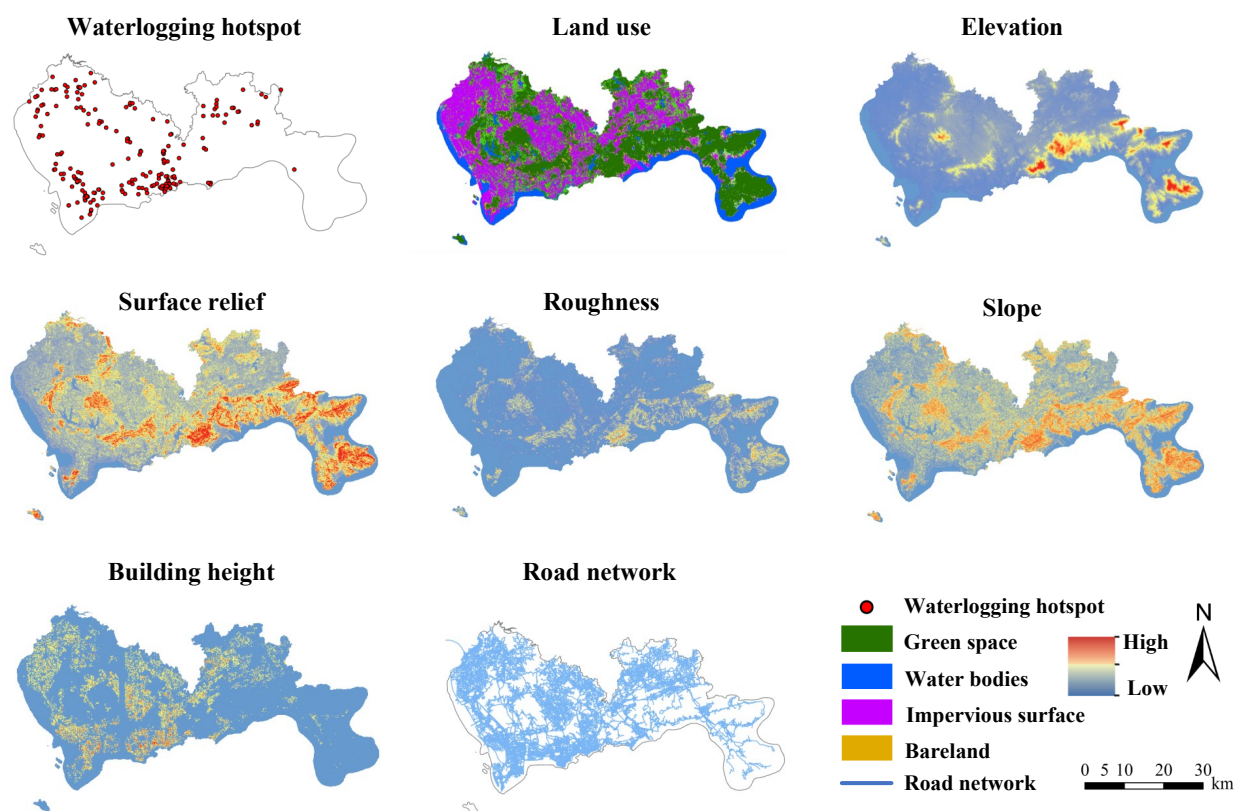


Figure 2. Waterlogging hotspots and potential drivers in Shenzhen.

2.2. Cost-Sensitive Framework for Dynamic Early Warning

This study formulates the prediction of waterlogging locations as a binary classification task, where waterlogging points are labeled as 1 and non-waterlogging points are labeled as 0. Therefore, the goal of waterlogging location prediction is to find a mapping function f :

$$f : R^D \rightarrow \{0, 1\} \quad (1)$$

where R^D represents the dimensionality of the input features. In this research, the input features comprise meteorological variables (e.g., rainfall, temperature, atmospheric pressure), land-use characteristics (e.g., impermeable surface ratio, green space ratio), and built-environment indicators (such as building density, road network density, etc.), amounting to a total of D features. Given a training dataset

$$(X, Y) \in \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}, \quad (2)$$

where $x_i \in \mathbb{R}^D$ and $y_i \in \{0, 1\}$, our objective is to identify the optimal mapping function that minimizes classification error.

As illustrated in Figure 3, we propose a comprehensive computational framework for urban waterlogging prediction. The framework proceeds through four sequential stages as follows:

(1) Hierarchical Feature Selection Strategy

To identify the minimal yet optimal feature subset, we employed a Lasso-guided iterative selection strategy. The Lasso method was specifically selected for its efficacy in managing multicollinearity among multi-source spatial variables, such as multi-scale precipitation and topographic factors. Lasso regularization was utilized to quantify the contribution of each feature, ranking them based on the magnitude of their non-zero coefficients. Unlike model-dependent importance rankings which may be biased by feature redundancy, Lasso provides clearer physical interpretability by assigning zero coefficients to redundant variables through L_1 regularization, effectively highlighting the most influential and independent predictors. Feature subsets of increasing sizes were then iteratively evaluated according to this ranking to analyze the trade-off between dimensionality and predictive accuracy. The final feature set was determined by identifying the performance saturation point where model stability is achieved to maximize parsimony without compromising detection sensitivity.

(2) Data balancing using SMOTE-WDC

To overcome the limitations of traditional oversampling, specifically the neglect of feature relevance and the generation of boundary noise, we implemented the Synthetic Minority Over-sampling Technique based on Weighted Distance and Cleaning (SMOTE-WDC) algorithm. This method optimizes the oversampling process through a hierarchical framework encompassing weighted metric construction, source noise elimination, and constrained sample generation. To precisely measure the similarity between samples containing variables of differing physical significance, we introduced a feature importance weighting mechanism. The algorithm utilizes Normalized Mutual Information (NMI) to quantify the correlation between each feature f_i and the target variable Y , where the feature weight w_i is computed as:

$$w_i = \frac{\text{NMI}(f_i, Y)}{\sum_{j=1}^M \text{NMI}(f_j, Y)} \quad (3)$$

Incorporating these weights with range normalization, the Weighted Gower Distance between sample x_a and x_b is defined as:

$$D_{WG}(x_a, x_b) = \sum_{k=1}^M w_k \frac{|x_{a,k} - x_{b,k}|}{R_k} \quad (4)$$

where R_k represents the range of the k -th feature. This reformulation ensures that the neighbor selection process aligns with the physical logic of the meteorological data rather than being dominated by numerical magnitudes.

Building upon this weighted metric, the algorithm incorporates a noise elimination step prior to oversampling. We employ the DBSCAN algorithm to identify and remove noise samples from the minority class, utilizing the proposed Weighted Gower Distance (D_{WG}) as the clustering metric. By filtering out outliers in the weighted feature space before interpolation, this step effectively prevents the propagation and amplification of low-quality samples, ensuring that synthetic data is derived from reliable seed samples.

To ensure the reliability of the synthetic data, a geometric Reverse Check Mechanism is applied during the generation phase. After a new sample x_{syn} is generated via linear interpolation, the algorithm calculates its Euclidean distance to the nearest minority instance (d_{min}) and the nearest majority instance (d_{maj}). The synthetic sample is retained only if it satisfies the constraint:

$$d_{min} < d_{maj} \quad (5)$$

This geometric constraint ensures that generated samples remain spatially closer to the waterlogging class, effectively filtering out ambiguous samples that encroach upon the majority class boundary and thereby enhancing the robustness of the classification boundary. The overall execution logic and iterative steps of the proposed SMOTE-WDC algorithm are summarized in Algorithm 1.

Algorithm 1 SMOTE-WDC Algorithm

Require: Minority class set X_{min} ; Majority class set X_{maj} ; Neighbors k ; Sampling ratio σ

Ensure: Augmented balanced dataset X_{new}

- 1: Initialize feature weights W using Normalized Mutual Information (NMI)
- 2: Construct Weighted Gower Distance matrix D for all $x \in X_{min}$ using W
- 3: **Preprocessing (DBSCAN-based Cleaning):**
- 4: Apply DBSCAN on X_{min} using distance matrix D
- 5: Remove noise samples to obtain filtered set X_{clean}
- 6: **Synthetic Generation and Post-cleaning:**
- 7: **while** $count < N$ **do**
- 8: Select $x_i \in X_{clean}$
- 9: Find k -nearest neighbors x_{nn} via D
- 10: Generate synthetic sample:

$$x_{syn} = x_i + rand(0, 1) \times (x_{nn} - x_i)$$

- 11: **Reverse Neighborhood Cleaning:**
 - 12: Compute $d_{min} = dist(x_{syn}, X_{clean})$
 - 13: Compute $d_{maj} = dist(x_{syn}, X_{maj})$
 - 14: **if** $d_{min} < d_{maj}$ **then**
 - 15: Add x_{syn} to X_{syn}
 - 16: $count \leftarrow count + 1$
 - 17: **end if**
 - 18: **end while**
 - 19: **return** $X_{new} = X_{maj} \cup X_{min} \cup X_{syn}$
-

(3) Model Comparative Optimization and Hyperparameter Tuning

To identify the optimal model for waterlogging prediction, we conducted a comprehensive crosswise comparison incorporating both ensemble architectures and non-ensemble benchmarks. We selected three representative tree-based ensemble learning algorithms—Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Extreme Gradient Boosting (XGBoost)—due to their proven ability to handle complex, non-linear spatial relationships. Furthermore, to ensure a robust evaluation and justify the necessity of ensemble methods, four diverse non-ensemble models, including Support Vector Machine (SVM-RBF), Multi-Layer Perceptron (MLP_NeuralNet), Logistic Regression, and K-Nearest Neighbors (KNN), were implemented as baseline comparators. This multi-model framework allows for a systematic assessment of predictive stability and accuracy across varying feature dimensions.

To identify the optimal model and its corresponding feature configuration, we conducted a comprehensive sensitivity analysis based on a Lasso-guided incremental feature selection strategy. Instead of treating features in isolation, we iteratively expanded the feature subset (ranging from 1 to 37 variables) according to their Lasso importance rankings to evaluate the sensitivity of model performance to different feature combinations. The final feature combination was determined by locating the performance saturation point. We integrated 5-fold cross-validation with grid search during the training phase to systematically explore the optimal hyperparameter combinations (including the number of trees, tree depth, and learning rate) within a predefined parameter space, aiming to maximize model performance on the validation set. The final model performance was then tested on the remaining two independent events (1 May and 26 July 2024), which were not involved in the training or tuning process. This strict separation provides a direct assessment of the model's reliability in predicting unseen waterlogging scenarios.

(4) Objective Function Setting and Threshold Calibration

To reconcile predictive robustness with the high-sensitivity requirements of disaster early warning, a multi-stage optimization strategy is implemented. The core model structure and hyperparameters are first established by maximizing the F1-score, which ensures an optimal equilibrium between Precision and Recall during the initial training phase. However, recognizing that the socio-economic cost of missed warnings significantly outweighs that of false alarms, the optimization target shifts to the $F\beta$ -score, specifically setting $\beta = 2$. This configuration assigns higher importance to Recall, thereby imposing a more rigorous penalty on false negatives during the validation phase. The calibration is executed by evaluating the F2-score across varying probability thresholds on the validation subsets. For each cross-validation fold, the threshold that maximizes the F2 metric is identified as the local optimal. To ensure the stability of the early warning system, the final operational threshold is determined by calculating the mean of these optimal thresholds across all folds. This systematic calibration ensures that the framework prioritized the capture of hazard events, providing a safety-first decision mechanism for urban emergency management.

Through the comprehensive application of the aforementioned methods, we constructed a practical computational framework for urban waterlogging location prediction, providing robust support for the precise prevention and control of urban waterlogging. In the subsequent results analysis section, we will systematically validate the predictive efficacy and practical application effects of the proposed framework, alongside in-depth analysis of key driving factors shaping urban waterlogging patterns.

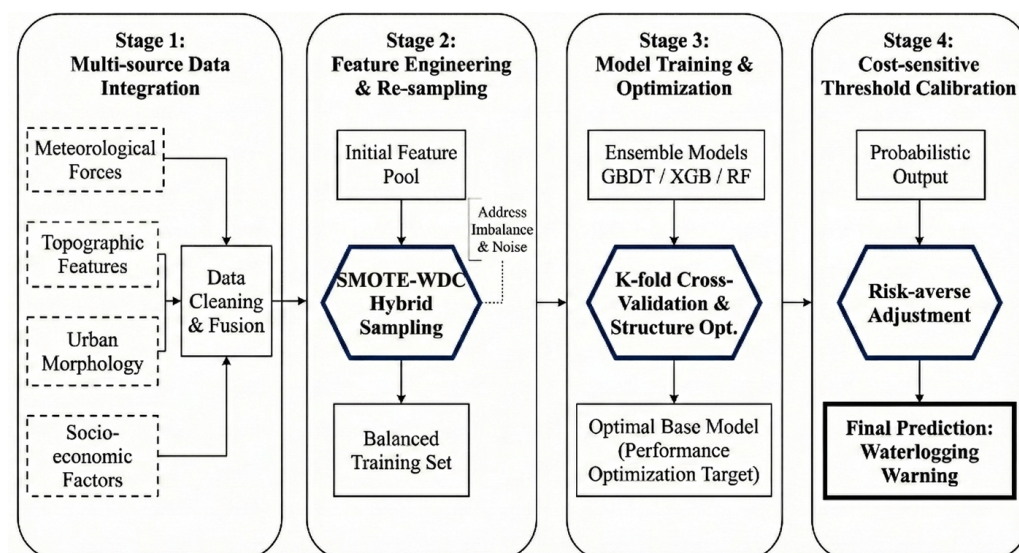


Figure 3. The hierarchical framework for urban waterlogging prediction.

2.3. Performance Evaluation

To rigorously evaluate the framework's efficacy in identifying sparse waterlogging hotspots, we adopt a multi-dimensional evaluation strategy tailored to severe class imbalance. The Area Under the Precision-Recall Curve (AUC-PR) is prioritized as the primary criterion for model selection, as it provides a robust assessment of predictive performance across all potential decision thresholds. While Recall directly embodies the safety margin of the early warning system by quantifying the capture of actual events, Precision ensures the reliability of positive predictions. The F1-score effectively harmonizes these two conflicting objectives, providing a measure of the model's structural robustness.

To further correct for agreement that could occur merely by chance, we incorporate Cohen's Kappa coefficient κ to assess statistical consistency:

$$\kappa = (p_0 - p_e) / (1 - p_e) \quad (6)$$

where p_0 denotes observed agreement and p_e represents the probability of random agreement. A high Kappa value confirms the model's capacity to capture meaningful spatiotemporal patterns rather than relying on class distribution bias.

Together, these metrics provide a holistic assessment of the model across critical dimensions, offering a comprehensive basis for model optimization and selection while ensuring its predictive capability and reliability in practical early warning applications.

3. Results

3.1. Prediction Results

To identify the optimal configuration that yields both high Recall and Precision, prioritizing AUC-PR as the primary selection criterion, we conducted a systematic evaluation incorporating hyperparameter grid search for each feature subset size ranging from 1 to 37 variables. This evaluation encompassed both ensemble architectures (GBDT, XGBoost, and Random Forest) and diverse non-ensemble benchmarks (SVM-RBF, MLP_NeuralNet, Logistic Regression, and KNN). As illustrated in Figure 4, the GBDT model achieved the global maximum performance at the 5-feature threshold, representing the most effective coupling between a parsimonious set of spatial predictors and urban waterlogging occurrences.

The observed fluctuations and subsequent performance decline beyond 5 features provide critical evidence of feature redundancy. In the high-dimensional space of ur-

ban datasets, the inclusion of lower-ranked variables—characterized by higher multicollinearity and weaker physical correlation—introduced marginal noise into the decision trees. This is particularly evident in the Random Forest model, which exhibited significant sensitivity to feature redundancy.

The predictive efficacy of the proposed framework was systematically evaluated through a comparative analysis of seven different machine learning algorithms across varying feature dimensions. The performance metrics of the 15 most effective models, ranked by their AUC-PR coefficients, are summarized in Table 2. Notably, the GBDT model demonstrated exceptional robustness by securing the top 8 positions, achieving its global peak performance with a condensed set of 5 core features (GBDT-5). This optimal configuration yielded a maximum AUC-PR of 0.895, a Recall of 0.882, and a Kappa coefficient of 0.776. To further justify this selection, Figure 5 provides a horizontal comparison of the predictive performance across the three ensemble models specifically at this optimal 5-feature threshold. While Random Forest achieves a marginally higher Recall of 0.93, it suffers from a significantly lower Precision of 0.52, leading to excessive false alarms. Similarly, XGBoost matches GBDT's high sensitivity with a Recall of 0.88 but falls short in prediction reliability with a lower Precision of 0.66. GBDT provides the most robust equilibrium for urban emergency management, effectively overcoming the excessive missed detections (false negatives) observed in the unoptimized baseline while maintaining the highest discriminative.

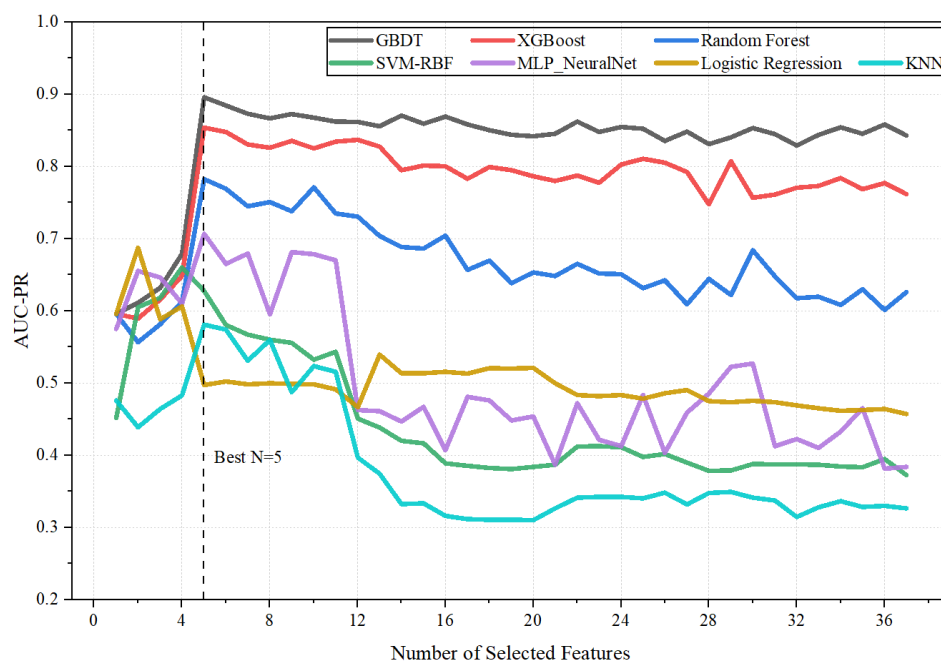


Figure 4. Model performance comparison across feature scales.

Two key technical strategies underpin this high performance. First, the integration of the SMOTE-WDC algorithm effectively refined minority class boundaries, ensuring sustained stability in the sparse feature space and preventing the drastic fluctuations observed in other ensemble variants. Second, the implementation of the F2-score as the optimization target played a pivotal role in cost-sensitive threshold calibration. This strategy prioritized Recall to meet the high-sensitivity requirements of urban emergency management, allowing GBDT-5 to capture 88.18% of actual inundation events while maintaining a reliable Precision of 0.746. As illustrated by the AUC-PR trends in Figure 6, the GBDT framework consistently achieves the highest discriminative capacity, providing a scientifically grounded tool for real-time waterlogging prediction.

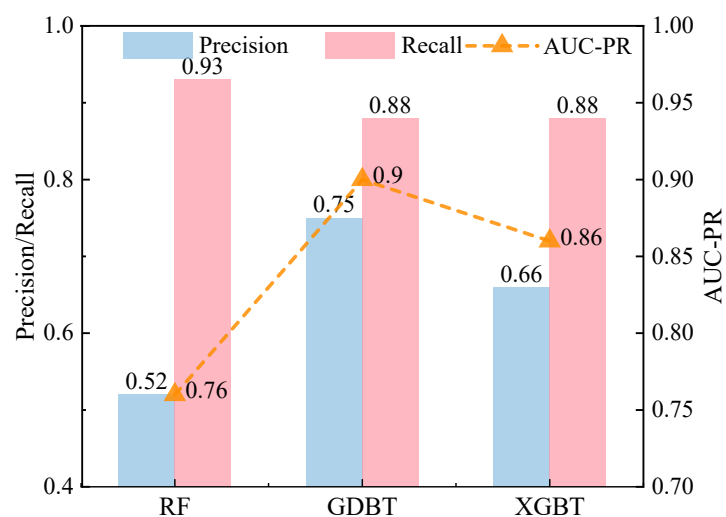


Figure 5. Performance metrics analysis of waterlogging prediction models.

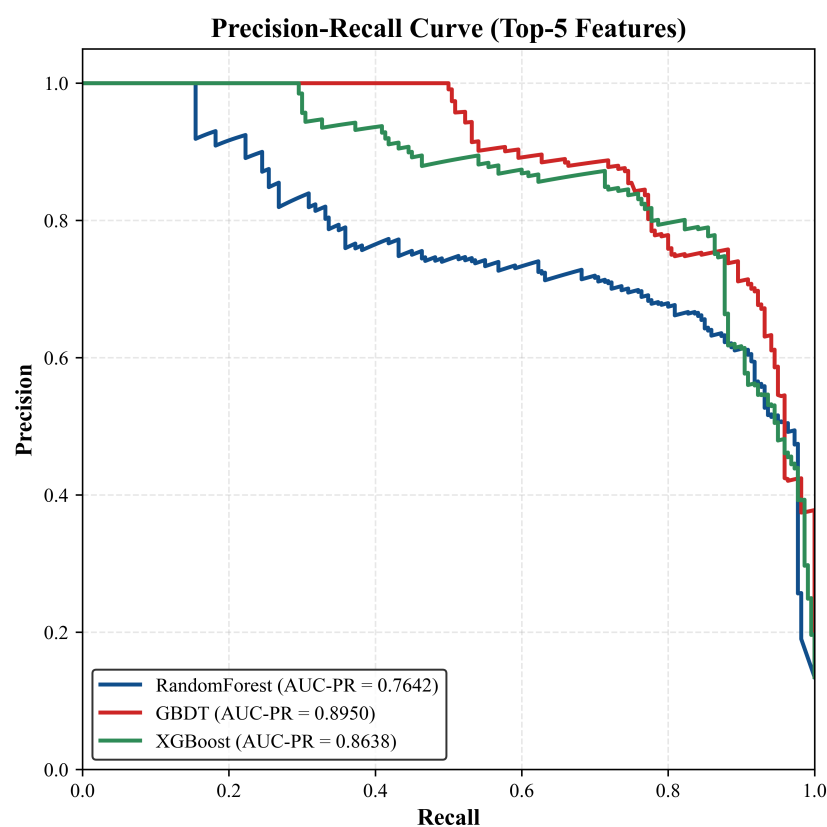


Figure 6. Model performance evaluation using AUC-PR.

Table 2. Performance comparison of the top 15 model configurations (Ranked by AUC-PR).

Rank	Feature Number	Model	Precision	Recall	F1-Score	AUC-PR	κ
1	5	GBDT	0.746	0.882	0.808	0.895	0.776
2	6	GBDT	0.704	0.832	0.763	0.884	0.722
3	10	GBDT	0.672	0.896	0.768	0.872	0.726
4	7	GBDT	0.712	0.832	0.767	0.873	0.728
5	9	GBDT	0.679	0.886	0.769	0.872	0.728
6	13	GBDT	0.673	0.859	0.755	0.871	0.711
7	14	GBDT	0.664	0.873	0.754	0.869	0.710
8	8	GBDT	0.686	0.905	0.780	0.866	0.741

Table 2. Cont.

Rank	Feature Number	Model	Precision	Recall	F1-Score	AUC-PR	κ
9	5	XGBoost	0.610	0.882	0.755	0.864	0.710
10	11	GBDT	0.654	0.886	0.753	0.863	0.708
11	12	GBDT	0.681	0.846	0.755	0.862	0.712
12	18	GBDT	0.674	0.846	0.750	0.861	0.706
13	7	XGBoost	0.668	0.914	0.772	0.858	0.730
14	9	XGBoost	0.643	0.868	0.739	0.855	0.691
15	16	GBDT	0.677	0.868	0.761	0.855	0.718

3.2. Feature Importance Analysis

To gain deeper insight into the driving mechanisms of urban waterlogging and validate the physical consistency of the GBDT model, we analyzed the contribution of the five selected features using both global SHAP summary analysis and local waterfall decomposition (Figure 7). The results reveal a clear hierarchical influence dominated by historical susceptibility and multi-dimensional meteorological forcing on prediction outcomes.

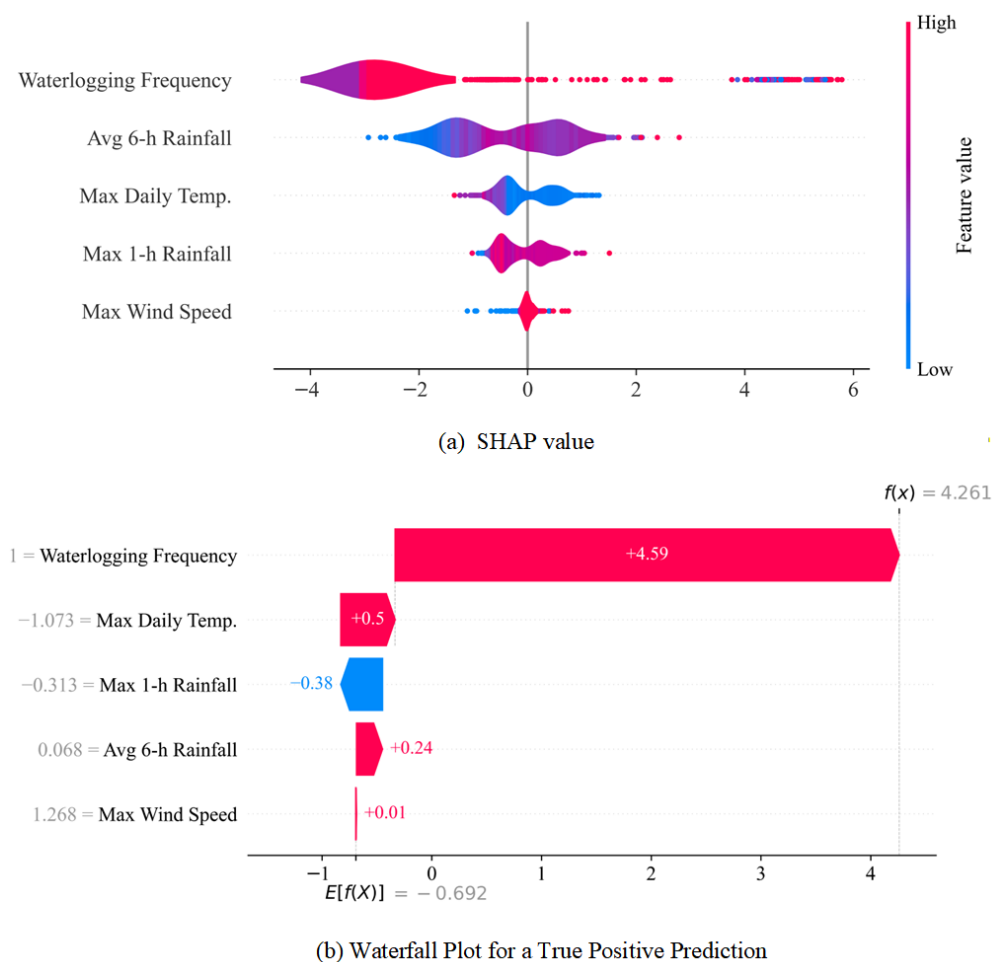


Figure 7. Interpretability analysis of the GBDT model using SHAP values.

The analysis identifies Waterlogging Frequency as the most dominant predictor, exhibiting the highest SHAP value and a strong positive correlation with waterlogging risk. From a physical perspective, this feature acts as a comprehensive proxy for intrinsic vulnerability, effectively capturing unmeasured static constraints such as topographical depressions and aging drainage infrastructure. The model's heavy reliance on historical priors

indicates that it successfully utilizes site-specific vulnerability to calibrate risk probabilities, prioritizing areas with known drainage deficiencies.

In terms of meteorological triggers, the model demonstrates a sophisticated understanding of hydrological dynamics by capturing both intensity and cumulative effects. Average 6 h rainfall (Avg 6 h Rainfall) and Maximum 1 h rainfall (Max 1 h Rainfall) are identified as the primary hydrometeorological drivers. The former reflects the potential for saturation-excess runoff caused by sustained precipitation, while the latter captures the immediate impact of high-intensity flash floods. Additionally, Maximum Daily Temperature (Max Daily Temp.) exhibits a notable influence, representing atmospheric coupling effects where higher temperatures correlate with increased convective energy and extreme precipitation potential within urban heat islands. Maximum Wind Speed (Max Wind Speed) further contributes to the decision logic, likely reflecting the impact of wind-driven rain on drainage efficiency or the association between high wind speeds and the passage of intense storm fronts.

To further illustrate the decision logic at the instance level, a waterfall plot of a selected True Positive (TP) prediction is presented in Figure 7b. This local explanation shows how the base value (expected model output) is adjusted by each feature to arrive at the final prediction ($f(x) = 4.261$). In this specific instance, the high Waterlogging Frequency provides a substantial positive push (+4.59), confirming the site's high intrinsic vulnerability. Although the Max 1 h Rainfall in this case shows a minor negative contribution (−0.38) relative to the average, the synergistic effects of Max Daily Temp (+0.5) and Avg 6 h Rainfall (+0.24) collectively reinforce the final waterlogging prediction. This decomposition confirms the model's ability to balance historical priors with real-time meteorological fluctuations to reach a reliable decision.

In summary, the combined SHAP summary and waterfall analysis confirms that the GBDT model has moved beyond simple pattern matching to capture the synergistic effects of historical vulnerability and multi-variable meteorological forcing. This validates that the proposed feature selection strategy successfully retained a parsimonious set of predictors with strong physical interpretability.

3.3. Ablation Study: Effectiveness of Proposed Strategies

To rigorously quantify the contribution of each proposed module and ensure experimental transparency, a stepwise ablation study was conducted across five scenarios (S1–S5). To maintain strict fairness and isolate the independent impact of each strategy, all experimental variants were executed under a controlled environment using a unified random seed, a predefined set of GBDT hyperparameters ($n_estimators = 200$, $max_depth = 4$, and $learning_rate = 0.05$) and F2-score based threshold optimization. Notably, Lasso-based feature importance ranking was pre-applied to the entire dataset to ensure a consistent prioritization of predictors, and a stratified 5-fold cross-validation scheme was consistently employed to maintain stable and reproducible evaluation across all scenarios. The Scenario S1 (Baseline), utilizing the complete 37-feature set and a default decision to establish the raw performance benchmark. Scenario S2 transitions to the 5 Lasso-selected features. S3 incorporates Standard Synthetic Minority Over-sampling Technique (Standard SMOTE) to provide a benchmark for traditional oversampling under cost-sensitive conditions. Scenario S4 introduces the SMOTE based on Weighted Distance (SMOTE-WD) algorithm, which utilizes a Weighted Gower Distance (WD) derived from Normalized Mutual Information (NMI) to guide synthetic sample generation, ensuring the physical consistency of the augmented data. Finally, Scenario S5 (SMOTE-WDC) represents the full framework, integrating the weighted distance mechanism with a DBSCAN-based cleaning phase to

systematically remove boundary noise and synthetic artifacts. The experimental results across five scenarios are summarized in Table 3.

Table 3. Stepwise ablation analysis of the proposed framework components.

Scenario	Methodology	Precision	Recall	F1-Score	AUC-PR	Kappa
S1	Baseline (Raw Data)	0.807	0.836	0.821	0.895	0.793
S2	+Feature Selection (FS)	0.862	0.796	0.827	0.929	0.802
S3	+FS + Standard SMOTE	0.759	0.832	0.794	0.914	0.760
S4	+FS + SMOTE-WD	0.767	0.855	0.809	0.897	0.777
S5	+FS + SMOTE-WDC (Proposed)	0.746	0.882	0.808	0.895	0.776

The ablation results validate the necessity of each module in addressing the challenges of feature redundancy and class imbalance (Figure 8). The analysis begins with the impact of feature reduction. Comparing the Baseline (S1) with S2, the introduction of the Lasso-based feature selection strategy yielded the most significant improvement in overall model quality, with the AUC-PR increasing from 0.895 to 0.929 and the F1-score rising to 0.827. This confirms that raw multi-source datasets contain substantial noise and spurious correlations; by distilling the variables, the model significantly reduced overfitting and established a highly reliable predictive baseline.

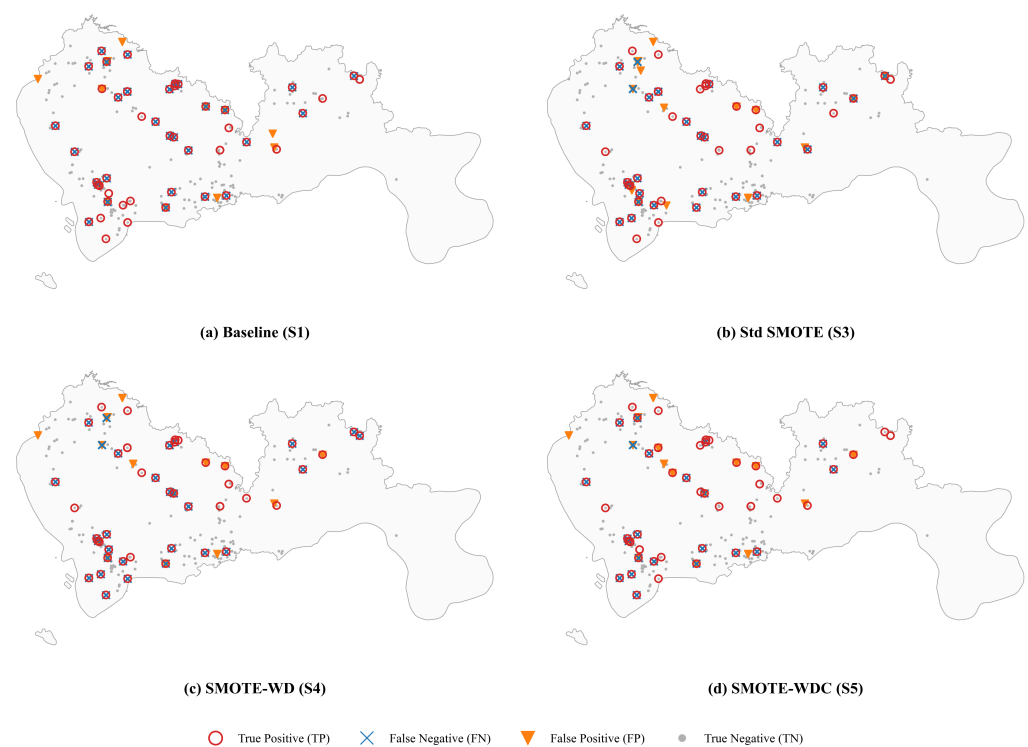


Figure 8. Visual comparison of different oversampling strategies for identifying waterlogging hotspots.

To satisfy the high-sensitivity requirements of urban emergency response, where minimizing missed warnings is paramount, we evaluated different oversampling strategies. As shown in Scenario 3, applying Standard SMOTE actually led to a rise in Recall (0.832) compared to S2. The SMOTE-WD algorithm (S4) effectively addressed the class imbalance by generating minority samples that respect the unique weights of different environmental predictors, thereby pushing the Recall to 0.855.

The final scenario, SMOTE-WDC (S5), introduces instance-level noise cleaning to the SMOTE-WD foundation. By removing synthetic samples that blur the decision boundaries, S5 achieved the highest Recall of 0.882, representing a decisive 4.6% improvement over

the Baseline (S1). Although this gain in sensitivity incurred a calculated trade-off in Precision (0.746), it is a necessary optimization to meet the “zero-miss” objective of practical waterlogging early warning.

To further validate the efficacy of the proposed framework, we conducted a comprehensive comparative analysis between the SMOTE-WDC algorithm and several representative resampling variants. These benchmarks include Standard Synthetic Minority Over-sampling Technique (Standard SMOTE), Borderline-SMOTE, Synthetic Minority Over-sampling Technique combined with Tomek Links (SMOTE-Tomek), and Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors (SMOTE-ENN). As summarized in Table 4, the SMOTE-WDC model demonstrates a superior capacity for capturing critical waterlogging signals. Notably, it achieves the highest Recall (0.882) among all tested methods. While Borderline-SMOTE exhibits a higher Precision (0.838), its significantly lower Recall (0.777) indicates a high rate of missed detections. In contrast, our proposed method maintains a robust Kappa coefficient (0.776)—ranking second only to Borderline-SMOTE (0.778)—which underscores its ability to balance class identification with overall predictive reliability. To ensure a fair comparison and isolate the independent contribution of each algorithm, all experiments utilized identical hyperparameters and experimental settings as the ablation study, including the 5-fold cross-validation, and F2-score based threshold optimization.

The performance gain is primarily attributed to the dual-stage refinement mechanism of SMOTE-WDC. By integrating Weighted Distance (WD) based on historical vulnerability, the model ensures that synthetic samples are generated within physically consistent bounds rather than through simple linear interpolation. Subsequently, the Density-based Cleaning (WDC) stage effectively prunes overlapping noise and “ghost” hotspots at the decision boundaries. This combination allows the model to prioritize high-sensitivity detection (Recall) while sharpening the distinction between flooded and non-flooded classes, providing a more reliable foundation for urban emergency response.

Table 4. Performance Comparison of Different Methodologies.

Methodology	Precision	Recall	AUC-PR	Kappa
SMOTE-WDC	0.746	0.882	0.895	0.776
Standard SMOTE	0.759	0.832	0.914	0.760
SMOTE-Tomek	0.736	0.823	0.902	0.740
SMOTE-ENN	0.698	0.800	0.865	0.703
Borderline-SMOTE	0.838	0.777	0.904	0.778

In summary, the ablation study confirms that Feature Selection ensures the model’s fundamental accuracy, while SMOTE-WDC serves as the critical engine for sensitivity enhancement, enabling the high recall necessary for practical urban waterlogging early warning.

4. Conclusions

To address the limitations of existing studies regarding feature redundancy, boundary noise in imbalanced data, and the lack of physical interpretability, this study proposes an interpretable machine learning prediction framework for urban waterlogging. By integrating physics-informed feature selection, the SMOTE-WDC sampling strategy, and cost-sensitive optimization, the framework achieves the accurate identification of waterlogging risk points and meets the stringent requirements for emergency response.

The key findings are as follows:

- (1) **Optimization of Sparse Feature Space and Physical Consistency:** We constructed a comprehensive urban waterlogging dataset integrating dynamic meteorology, topography, vertical urban morphology, and demographic distribution. A critical finding of this study is the successful identification of a parsimonious 5-feature subset that yielded the global peak in predictive performance (AUC-PR 0.895, Kappa 0.776). This selection process effectively filtered out high-dimensional noise and redundant correlations, distilling the complex feature space into its most critical physical drivers. SHAP value analysis validated this selection by quantifying the dominant contribution of historical vulnerability (Waterlogging Frequency) and identifying the synergistic roles of short-term intense rainfall, sustained accumulation, and atmospheric coupling effects (temperature and wind speed). This transition from a high-dimensional baseline to a high-performance sparse model confirms that the framework successfully captures the core features governing urban inundation.
- (2) **Effectiveness of SMOTE-WDC in Noise Mitigation:** A comprehensive comparative analysis against several well-established resampling algorithms (e.g., Borderline-SMOTE, SMOTE-Tomek, and SMOTE-ENN) verified the superiority of our proposed SMOTE-WDC. While traditional approaches often degraded precision by introducing boundary noise, SMOTE-WDC effectively refined the feature space. By utilizing Weighted Gower Distance for generation and DBSCAN for cleaning, it successfully enriched minority sample. This provided a more robust structural foundation for identifying complex waterlogging events than standard incremental approaches.
- (3) **High Recall for Real-time Early Warning:** To satisfy the practical demand for minimizing missed detections, the framework implemented a cost-sensitive threshold calibration. The operational feasibility of the model is supported by its low computational complexity and the high-frequency update cycle of Shenzhen's meteorological data. Given the sparse 5-feature input space, the trained model performs rapid inference with sub-second latency on a standard CPU. Since updated rainfall forecasts and dynamic triggers are available on a daily basis, the framework can perform rapid inference and generate forward-looking risk maps before the actual storm arrival. This optimization allowed the final model to achieve a Recall of 0.882, representing a decisive 4.6% relative improvement over the baseline. Although this involved a necessary trade-off in precision, it ensures the effective capture of actual waterlogging points, offering a reliable decision support tool that balances public safety with operational efficiency.

Despite these findings, certain practical constraints in model maintenance and spatial transferability warrant further consideration. Since the framework incorporates 'Historical Waterlogging Frequency' as a critical spatial prior, its predictive accuracy is tied to the urban infrastructure at the time of data collection. Significant modifications, such as the installation of new drainage systems or large-scale land-use changes, may introduce a temporary misalignment in predictions for those specific locations. To maintain long-term reliability, this framework is intended as an iterative learning system that should be periodically retrained with updated event data to internalize urban evolutions. It should be noted that the current validation is focused on Shenzhen, leveraging its unique and comprehensive waterlogging sensing infrastructure. The primary obstacle to cross-city expansion lies in the non-uniformity of sensing technologies and the restricted public access to high-frequency water depth records in most metropolitan areas. Despite these data-related constraints, the model's stability across multiple independent storm events in 2023 and 2024 provides a strong foundation for its reliability. Future work will explore domain adaptation techniques to bridge the gap between different urban sensing environments.

In summary, this study establishes a practical and interpretable approach for urban waterlogging prediction. By combining domain knowledge with advanced imbalanced learning techniques, the proposed framework allows urban planners and emergency managers to accurately identify high-risk areas and deploy resources effectively.

Furthermore, the study's implications extend beyond the specific case of Shenzhen. As a representative high-density coastal megacity, Shenzhen exemplifies the critical challenges of rapid urbanization, complex underlying surfaces, and intense monsoon precipitation. Since these characteristics are shared by many developing metropolises globally, the feature insights and the optimization framework derived in this study offer valuable guidance for waterlogging risk assessment in other cities facing similar urbanization pressures.

Looking ahead, future research will extend this framework to cities with diverse climatic and topographic conditions to validate its generalizability as standardized sensing networks and public data-sharing platforms become more widely available. Furthermore, we aim to incorporate micro-scale 3D urban features and employ causal inference techniques, thereby moving beyond prediction to uncover the deep-seated driving mechanisms of urban waterlogging.

Author Contributions: Conceptualization, Y.D. and X.L.; methodology, Y.D.; validation, Y.D.; investigation, Y.D.; resources, X.L.; data curation, Y.D.; writing—original draft preparation, Y.D.; writing—review and editing, Y.D.; visualization, Y.D.; supervision, X.L.; project administration, X.L.; funding acquisition, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (72025405, 72421002) and the Hunan Science and Technology Plan Project (2023JJ40685, 2024RC3133).

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

Acknowledgments: The authors would like to thank the Shenzhen Meteorological Bureau and relevant departments for providing the foundational data. We also appreciate the anonymous reviewers for their constructive comments.

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

References

1. Li, Q.; Sheng, B.; Huang, J.; Li, C.; Song, Z.; Chao, L.; Sun, W.; Yang, Y.; Jiao, B.; Guo, Z.; et al. Different climate response persistence causes warming trend unevenness at continental scales. *Nat. Clim. Change* **2022**, *12*, 343–349. [\[CrossRef\]](#)
2. Winsemius, H.C.; Aerts, J.C.; Van Beek, L.P.; Bierkens, M.F.; Bouwman, A.; Jongman, B.; Kwadijk, J.C.; Ligtoet, W.; Lucas, P.L.; Van Vuuren, D.P.; et al. Global drivers of future river flood risk. *Nat. Clim. Change* **2016**, *6*, 381–385. [\[CrossRef\]](#)
3. Kaiser, Z.R.M.A.; Akter, F. From risk to resilience and sustainability: Addressing urban flash floods and waterlogging. *Risk Sci.* **2025**, *1*, 100011. [\[CrossRef\]](#)
4. Chen, Y.; Hu, M.; Chen, X.; Wang, F.; Liu, B.; Huo, Z. An approach of using social media data to detect the real time spatio-temporal variations of urban waterlogging. *J. Hydrol.* **2023**, *625*, 130128. [\[CrossRef\]](#)
5. Peng, L.; Wang, Y.; Yang, L.; Garchagen, M.; Deng, X. A comparative analysis on flood risk assessment and management performances between Beijing and Munich. *Environ. Impact Assess. Rev.* **2024**, *104*, 107319. [\[CrossRef\]](#)
6. Shi, J.; Cui, L.; Tian, Z. Spatial and temporal distribution and trend in flood and drought disasters in East China. *Environ. Res.* **2020**, *185*, 109406. [\[CrossRef\]](#)
7. Su, M.; Zheng, Y.; Hao, Y.; Chen, Q.; Chen, S.; Chen, Z.; Xie, H. The influence of landscape pattern on the risk of urban water-logging and flood disaster. *Ecol. Indic.* **2018**, *92*, 133–140. [\[CrossRef\]](#)
8. Lu, X.; Wrathall, D.J.; Sundsoy, P.R.; Nadiruzzaman, M.; Wetter, E.; Iqbal, A.; Qureshi, T.; Tatem, A.J.; Canright, G.S.; Engo-Monsen, K.; et al. Detecting climate adaptation with mobile network data in Bangladesh: Anomalies in communication, mobility and consumption patterns during cyclone Mahasen. *Clim. Change* **2016**, *138*, 505–519. [\[CrossRef\]](#)

9. Lu, X.; Wrathall, D.J.; Sundsøy, P.R.; Nadiruzzaman, M.; Wetter, E.; Iqbal, A.; Qureshi, T.; Tatem, A.; Canright, G.; Engø-Monsen, K.; et al. Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh. *Glob. Environ. Change* **2016**, *38*, 1–7. [\[CrossRef\]](#)
10. Berndtsson, R.; Becker, P.; Persson, A.; Aspegren, H.; Haghighatafshar, S.; Jönsson, K.; Larsson, R.; Mobini, S.; Mottaghi, M.; Nilsson, J.; et al. Drivers of changing urban flood risk: A framework for action. *J. Environ. Manag.* **2019**, *240*, 47–56. [\[CrossRef\]](#)
11. Miller, J.D.; Hutchins, M. The impacts of urbanisation and climate change on urban flooding and urban water quality: A review of the evidence concerning the United Kingdom. *J. Hydrol. Reg. Stud.* **2017**, *12*, 345–362. [\[CrossRef\]](#)
12. Smith, B.K.; Smith, J.A.; Baeck, M.L.; Villarini, G.; Wright, D.B. Spectrum of storm event hydrologic response in urban watersheds. *Water Resour. Res.* **2013**, *49*, 2649–2663. [\[CrossRef\]](#)
13. Chao, L.; Zhang, K.; Yang, Z.L.; Wang, J.; Lin, P.; Liang, J.; Li, Z.; Gu, Z. Improving flood simulation capability of the WRF-Hydro-RAPID model using a multi-source precipitation merging method. *J. Hydrol.* **2021**, *592*, 125814. [\[CrossRef\]](#)
14. Tehrany, M.S.; Jones, S.; Shabani, F. Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. *Catena* **2019**, *175*, 174–192. [\[CrossRef\]](#)
15. Hunter, N.M.; Bates, P.D.; Horritt, M.S.; Wilson, M.D. Simple spatially-distributed models for predicting flood inundation: A review. *Geomorphology* **2007**, *90*, 208–225. [\[CrossRef\]](#)
16. Climate Change (IPCC). *Special Report: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*; Cambridge University Press: Cambridge, UK, 2012.
17. Forero-Ortiz, E.; Martínez-Gomariz, E.; Cañas Porcuna, M. A review of flood impact assessment approaches for underground infrastructures in urban areas: A focus on transport systems. *Hydrol. Sci. J.* **2020**, *65*, 1943–1955. [\[CrossRef\]](#)
18. Bruwier, M.; Maravat, C.; Mustafa, A.; Teller, J.; Piroton, M.; Erpicum, S.; Archambeau, P.; Dewals, B. Influence of urban forms on surface flow in urban pluvial flooding. *J. Hydrol.* **2020**, *582*, 124493. [\[CrossRef\]](#)
19. Li, J.; Bortolot, Z.J. Quantifying the impacts of land cover change on catchment-scale urban flooding by classifying aerial images. *J. Clean. Prod.* **2022**, *344*, 130992. [\[CrossRef\]](#)
20. Wang, L.; Li, Y.; Hou, H.; Chen, Y.; Fan, J.; Wang, P.; Hu, T. Analyzing spatial variance of urban waterlogging disaster at multiple scales based on a hydrological and hydrodynamic model. *Nat. Hazards* **2022**, *114*, 1915–1938. [\[CrossRef\]](#)
21. Du, W.; Gong, Y.; Chen, N. PSO-WELLSVM: An integrated method and its application in urban waterlogging susceptibility assessment in the central Wuhan, China. *Comput. Geosci.* **2022**, *161*, 105079. [\[CrossRef\]](#)
22. Yang, J.; Duan, C.; Wang, H.; Chen, B. Spatial supply-demand balance of green space in the context of urban waterlogging hazards and population agglomeration. *Resour. Conserv. Recycl.* **2023**, *188*, 106662. [\[CrossRef\]](#)
23. Li, H.; Wang, Q.; Li, M.; Zang, X.; Wang, Y. Identification of urban waterlogging indicators and risk assessment based on MaxEnt Model: A case study of Tianjin Downtown. *Ecol. Indic.* **2024**, *158*, 111354. [\[CrossRef\]](#)
24. Li, X.; Erpicum, S.; Mignot, E.; Archambeau, P.; Piroton, M.; Dewals, B. Influence of urban forms on long-duration urban flooding: Laboratory experiments and computational analysis. *J. Hydrol.* **2021**, *603*, 127034. [\[CrossRef\]](#)
25. Son, C.H.; Ban, Y.U. Flood vulnerability characteristics considering environmental justice and urban disaster prevention plan in Seoul, Korea. *Nat. Hazards* **2022**, *114*, 3185–3204. [\[CrossRef\]](#)
26. Li, Z.; Wang, Z.; Wu, L.; Huang, W.; Peng, W. Evaluating the effect of building patterns on urban flooding based on a boosted regression tree: A case study of Beijing, China. *Hydrol. Process.* **2023**, *37*, e14932. [\[CrossRef\]](#)
27. Vorogushyn, S.; Bates, P.D.; de Bruijn, K.; Castellarin, A.; Kreibich, H.; Priest, S.; Schröter, K.; Bagli, S.; Blöschl, G.; Domeneghetti, A.; et al. Evolutionary leap in large-scale flood risk assessment needed. *WIREs Water* **2017**, *5*, e1266. [\[CrossRef\]](#)
28. Yang, H.; Ning, W.; Wang, Z.; Sun, X. Determination and Sensitivity Analysis of Urban Waterlogging Driving Factors Based on Spatial Analysis Method. *Sustainability* **2025**, *17*, 2785. [\[CrossRef\]](#)
29. Kundu, S.; Lakshmi, V.; Torres, R. Flood Depth Estimation during Hurricane Harvey Using Sentinel-1 and UAVSAR Data. *Remote Sens.* **2022**, *14*, 1450. [\[CrossRef\]](#)
30. Unfer, G.R.; França, J.R.d.A.; Menezes, W.F.; Silva, F.P.d. A multiscale study of a heavy rainfall event of April 2019 in Rio de Janeiro city. *Nat. Hazards* **2024**, *121*, 7281–7308. [\[CrossRef\]](#)
31. Investigation Report on the “7·20” Extraordinary Rainstorm Disaster in Zhengzhou, Henan Province Released. Available online: <http://www.cfdm.cn/cn/article/id/2957> (accessed on 5 March 2026).
32. Liu, C.; Li, Y.; Li, J. Geographic information system-based assessment of mitigating flash-flood disaster from green roof systems. *Comput. Environ. Urban Syst.* **2017**, *64*, 321–331. [\[CrossRef\]](#)
33. Mignot, E.; Dewals, B. Hydraulic modelling of inland urban flooding: Recent advances. *J. Hydrol.* **2022**, *609*, 127763. [\[CrossRef\]](#)
34. Wang, M.; Zhang, Y.; Bakhshipour, A.E.; Liu, M.; Rao, Q.; Lu, Z. Designing coupled LID–GREI urban drainage systems: Resilience assessment and decision-making framework. *Sci. Total Environ.* **2022**, *834*, 155267. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Rafiei-Sardooi, E.; Azareh, A.; Choubin, B.; Mosavi, A.H.; Clague, J.J. Evaluating urban flood risk using hybrid method of TOPSIS and machine learning. *Int. J. Disaster Risk Reduct.* **2021**, *66*, 102614. [\[CrossRef\]](#)

36. Zhang, X.; Kang, A.; Lei, X.; Wang, H. Urban drainage efficiency evaluation and flood simulation using integrated SWMM and terrain structural analysis. *Sci. Total Environ.* **2024**, *957*, 177442. [\[CrossRef\]](#)
37. Lingkai, S.; Jian, W.; Wangpeng, X.; Zhaobo, G.; Wei, X.; Shi, Z. Risk Assessment of Urban Flooding Using InfoWorks ICM Model: A Case Study of Tangxun Lake Watershed in Wuhan. In *Environmental Science and Technology: Sustainable Development II*; Environmental Science and Engineering; Springer: Cham, Switzerland, 2024; Chapter 11, pp. 135–156. [\[CrossRef\]](#)
38. Lu, X.; Qin, W. Informatics in the Era of AI. *Innov. Inform.* **2025**, *1*, 100002. [\[CrossRef\]](#)
39. Li, Z.; Zhou, Z.; Wang, H.; Li, X.; Shi, X.; Xiao, J.; Yang, Z.; Sun, M.; Li, X.; Jia, H. Artificial intelligence-incorporated prediction for urban flooding processes in the past 20 years: A critical review. *Environ. Model. Softw.* **2025**, *192*, 106525. [\[CrossRef\]](#)
40. Refadah, S.S. Development in flood forecasting: A comprehensive review of complex and machine learning models. *Phys. Chem. Earth Parts A/B/C* **2025**, *139*, 103975. [\[CrossRef\]](#)
41. Kim, S.; Matsumi, Y.; Pan, S.; Mase, H. A real-time forecast model using artificial neural network for after-runner storm surges on the Tottori coast, Japan. *Ocean Eng.* **2016**, *122*, 44–53. [\[CrossRef\]](#)
42. Wang, Q.; Tang, M.; Shi, P. Depth prediction of urban waterlogging based on BiTCN-GRU modeling. *PLoS ONE* **2025**, *20*, e0321637. [\[CrossRef\]](#)
43. Mosavi, A.; Edalatifar, M. A Hybrid Neuro-Fuzzy Algorithm for Prediction of Reference Evapotranspiration. In *Recent Advances in Technology Research and Education*; Lecture Notes in Networks and Systems; Springer: Cham, Switzerland, 2019; Chapter 31, pp. 235–243. [\[CrossRef\]](#)
44. Dineva, A.; Varkonyi-Koczy, A.R.; Tar, J.K. Fuzzy expert system for automatic wavelet shrinkage procedure selection for noise suppression. In Proceedings of the IEEE 18th International Conference on Intelligent Engineering Systems INES 2014, Tihany, Hungary, 3–5 July 2014; IEEE: New York, NY, USA, 2014. [\[CrossRef\]](#)
45. Suykens, J.A.K.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [\[CrossRef\]](#)
46. Tang, X.; Hong, H.; Shu, Y.; Tang, H.; Li, J.; Liu, W. Urban waterlogging susceptibility assessment based on a PSO-SVM method using a novel repeatedly random sampling idea to select negative samples. *J. Hydrol.* **2019**, *576*, 583–595. [\[CrossRef\]](#)
47. Wu, Z.; Zhou, Y.; Wang, H.; Jiang, Z. Depth prediction of urban flood under different rainfall return periods based on deep learning and data warehouse. *Sci. Total Environ.* **2020**, *716*, 137077. [\[CrossRef\]](#) [\[PubMed\]](#)
48. An, Z.; Yu, X.; Wang, C.; Zhang, Y.; Song, C. Embodied intelligence: Recent advances and future perspectives. *Innov. Inform.* **2025**, *1*, 100008. [\[CrossRef\]](#)
49. Yalcin, E. Assessing the impact of topography and land cover data resolutions on two-dimensional HEC-RAS hydrodynamic model simulations for urban flood hazard analysis. *Nat. Hazards* **2020**, *101*, 995–1017. [\[CrossRef\]](#)
50. Zhang, Q.; Wu, Z.; Guo, G.; Zhang, H.; Tarolli, P. Explicit the urban waterlogging spatial variation and its driving factors: The stepwise cluster analysis model and hierarchical partitioning analysis approach. *Sci. Total Environ.* **2021**, *763*, 143041. [\[CrossRef\]](#)
51. Zhang, Q.; Wu, Z.; Zhang, H.; Dalla Fontana, G.; Tarolli, P. Identifying dominant factors of waterlogging events in metropolitan coastal cities: The case study of Guangzhou, China. *J. Environ. Manag.* **2020**, *271*, 110951. [\[CrossRef\]](#)
52. Quan, R.S. Rainstorm waterlogging risk assessment in central urban area of Shanghai based on multiple scenario simulation. *Nat. Hazards* **2014**, *73*, 1569–1585. [\[CrossRef\]](#)
53. Wu, J.; Sha, W.; Zhang, P.; Wang, Z. The spatial non-stationary effect of urban landscape pattern on urban waterlogging: A case study of Shenzhen City. *Sci. Rep.* **2020**, *10*, 7369. [\[CrossRef\]](#)
54. Lin, J.; He, X.; Lu, S.; Liu, D.; He, P. Investigating the influence of three-dimensional building configuration on urban pluvial flooding using random forest algorithm. *Environ. Res.* **2021**, *196*, 110438. [\[CrossRef\]](#)
55. Blöschl, G.; Hall, J.; Viglione, A.; Perdigao, R.A.P.; Parajka, J.; Merz, B.; Lun, D.; Arheimer, B.; Aronica, G.T.; Bilibashi, A.; et al. Changing climate both increases and decreases European river floods. *Nature* **2019**, *573*, 108–111. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Du, S.; Van Rompaey, A.; Shi, P.; Wang, J. A dual effect of urban expansion on flood risk in the Pearl River Delta (China) revealed by land-use scenarios and direct runoff simulation. *Nat. Hazards* **2015**, *77*, 111–128. [\[CrossRef\]](#)
57. Chen, W.; Wu, C.; Zhang, H.; Huang, G. Assessing the Impact of Climate Change on the Waterlogging Risk in Coastal Cities: A Case Study of Guangzhou, South China. *J. Hydrometeorol.* **2017**, *18*, 1549–1562. [\[CrossRef\]](#)
58. Quan, R.S.; Liu, M.; Lu, M.; Zhang, L.J.; Wang, J.J.; Xu, S.Y. Waterlogging risk assessment based on land use/cover change: A case study in Pudong New Area, Shanghai. *Environ. Earth Sci.* **2010**, *61*, 1113–1121. [\[CrossRef\]](#)
59. Mair, M.; Zischg, J.; Rauch, W.; Sitzenfrie, R. Where to Find Water Pipes and Sewers?—On the Correlation of Infrastructure Networks in the Urban Environment. *Water* **2017**, *9*, 146. [\[CrossRef\]](#)
60. Tran, D.; Xu, D.; Dang, V.; Alwah, A.A.Q. Predicting Urban Waterlogging Risks by Regression Models and Internet Open-Data Sources. *Water* **2020**, *12*, 879. [\[CrossRef\]](#)

61. Li, C.; Liu, M.; Hu, Y.; Wang, H.; Zhou, R.; Wu, W.; Wang, Y. Spatial distribution patterns and potential exposure risks of urban floods in Chinese megacities. *J. Hydrol.* **2022**, *610*, 127838. [[CrossRef](#)]
62. Lloyd, C.T.; Chamberlain, H.; Kerr, D.; Yetman, G.; Pistolesi, L.; Stevens, F.R.; Gaughan, A.E.; Nieves, J.J.; Hornby, G.; MacManus, K.; et al. Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. *Big Earth Data* **2019**, *3*, 108–139. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.