Enhancement of land subsidence prediction capabilities using machine learning and SHAP value analysis with Sentinel-1 InSAR Data

Heng Su
Chongqing University of Posts and Telecommunications
Tingting Xu
Chongqing University of Posts and Telecommunications
Xiancai Xion
xiongxc@cqu.edu.cn
Chongqing Institute of Surveying and Monitoring for Planning and Natural Resources
Aohua Tian
Chongqing University of Posts and Telecommunications

Research Article

Keywords: Land Subsidence, Machine Learning, SHAP, Features Optimization

Posted Date: March 22nd, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3926697/v1

Additional Declarations: No competing interests reported.

Enhancement of land subsidence prediction capabilities using machine learning and SHAP value analysis with Sentinel-1 InSAR Data

Heng Su[a], Tingting Xu[a], Xiancai Xion[b,*], Aohua Tian[a]

[a]*School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China;*

[b]*Chongqing Institute of Surveying and Monitoring for Planning and Natural Resources, Chongqing, China;*

*Correspondence Author: xiongxc@cqu.edu.cn

Abstract： Land subsidence has been a significant focus of geoscience studies, and researching the factors contributing to it and predicting future incidents is essential. However, current research requires a systematic and coherent strategy to identify symptoms of land sinking with scientific rigor. This study employs neural networks and SHAP values to forecast land subsidence. We utilized SHAP values as a substitute for the usual random forest (RF) technique to evaluate the attributes of land subsidence. In addition, we employed neural networks to predict the prospective areas where land subsidence may occur in Chongqing and Chengdu in the future, considering a diverse set of conceivable situations. The results indicate that using neural networks for land subsidence prediction improves the model's accuracy by 16% compared to the traditional approach. The performance is enhanced by approximately 22% after optimizing the input characteristics. The feature optimization strategy proposed in this study, which relies on SHAP values, is particularly advantageous for predicting land subsidence. Moreover, the determinants of land sinking, ascertained by data analysis considering intricate topography, correspond with the conclusions of previous research. Utilizing this feature optimization method can improve the selection of input variables for the land subsidence prediction model, increasing accuracy and providing a solid theoretical foundation for preventing urban land subsidence.

Keywords: Land Subsidence; Machine Learning; SHAP; Features Optimization

1 Introduction

Land subsidence refers to the vertical sinking or downward movement of rock layers on or beneath the Earth's surface. Typically, it is linked to human actions or natural phenomena and has diverse effects on land utilization, infrastructure, and the environment. The origins of land subsidence can be traced back to ancient times, but its issues have become more apparent and intricate with the advent of modern industrialization and urbanization. Land subsidence can occur due to a variety of factors, such as the extraction of groundwater and minerals, seismic activity, and surface loads (Larson et al., 2001; Kumar et al., 2022; Wu et al., 2020; Yu et al., 2020). Land sinking has been detected in prominent urban areas, such as Beijing and Shanghai (Gu et al., 2014; He et al., 2018), throughout the past few years. Land subsidence causes significant damage to structures such as houses, roads, and underground pipes (Ohenhen et al., 2022; Park et al., 2022). Additionally, it results in surface drainage issues, alterations in groundwater quality, and harm to natural ecosystems (Saber et al., 2018; Zou et al., 2015). Hence, it is crucial to possess a precise comprehension of the circumstances and factors leading to land subsidence and employ suitable strategies to handle and alleviate its consequences to ensure sustainable land utilization and urban development. Researchers have recently conducted studies examining ground subsidence monitoring using different methodologies. Chen et al. (2016) established a correlation between land subsidence and groundwater level, active faults, and other parameters. They achieved this by analyzing Envisat and TerraSAR-X datasets and comparing the subsidence rates calculated using InSAR and GPS. Zhou et al. (2020) derived the results by comparing the subsidence characteristic areas of different land use types. Their findings indicate that ground subsidence is particularly pronounced in areas where land use is intricate and involves a combination of residential, industrial, and agricultural activities. Guo et al. (2021) employed a combination of remote sensing, physical exploration, spatial analysis, hydrogeology, and other multidisciplinary approaches to investigate the causes of abrupt ground subsidence on the Beijing Plain at various levels of analysis. These studies have examined the causes of land subsidence and associated factors and have determined that groundwater content is the primary factor influencing land subsidence. Nevertheless, the land settlement cannot be accurately predicted by a single element alone, as the collective impact of multiple traits influences it.

To construct a land settlement prediction model, it is essential to evaluate various criteria.

The majority of contemporary land subsidence prediction fields employ tree models. For example, Li et al. (2023) used four machine learning approaches using support vector machine (SVM), gradient boosted decision tree (GBDT), random forest (RF), and extreme random tree (ERT) to create surface subsidence rate prediction model and surface subsidence gradient prediction model. Ebrahimy et al. (2020) employed augmented regression tree (BRT), Random Forest (RF), and Classification and Regression Tree (CART) techniques to generate and compare Land Surface Sedimentation Sensitivity Maps (LSSM). They confirmed that the BRT model had a higher prediction accuracy (AUROC=0.819) compared to the RF model (AUROC=0.798) and CART model (AUROC=0.764). However, studies have yet to be undertaken to anticipate where land subsidence is likely to occur and to bring neural networks into land subsidence prediction. Neural networks may represent these complicated interactions as a versatile model through several hidden layers and nonlinear activation functions. In contrast, tree models are basic decision rules based on a partitioned data space and cannot handle nonlinear interactions. Wang et al. (2023) studied and predicted regional ground subsidence based on BP neural network (BPNN) and random forest (RF) approaches. Ma et al. (2020) introduced a data-driven strategy based on deep convolutional neural networks (DCNN) to predict land subsidence based on ground settlement based on interferometric synthetic aperture radar (InSAR) data. They proved that DCNN is effective for the short-term prediction of InSAR time series deformation. Zhou et al. (2022) projected urban ground settlement sequence deformation variations from a multi-factor perspective after employing the LSTM neural network technique for urban ground settlement prediction. Most of these researches employed several features to train the model and used applicable methods to verify its association with land sinking. However, only some tested if the neural network model is superior to the tree model in land subsidence prediction. The present study on land subsidence prediction demands a systematic set of feature selection techniques that can make the land subsidence prediction model more explanatory. Land subsidence is a complex process, and various contributing elements may contribute to land subsidence, such as groundwater, development of subways, massive structures, and rainfall in the area. However, in earlier studies, only some have taken an integrated perspective of these aspects, and no set of characteristic optimization approaches exist. Shi et al. (2020) used groundwater level variation, Quaternary sediment thickness, and exponential accumulation index (IBI) as influencing elements to forecast the extent of subsidence in the Beijing Plain. Still, they did not apply a scientific procedure to pick the land subsidence features. Li et al. (2022) identified the long-term over-exploitation of groundwater in the intermediate and deep aquifers and the mining of underground mineral resources as the key triggering drivers of ground subsidence in the SDP. However, the topic was only considered at a vast geographic scale.

Meanwhile, only limited research assesses the performance of land subsidence prediction between different models. Rafiei Sardooi et al. (2022) employed support vector machine (SVM) and random forest (RF) models for land subsidence prediction, analyzed the performance of these models using AUC values and TSS metrics, and validated that the prediction accuracy of SVM models was greater than that of RF models. However, feature screening or evaluation of land subsidence indicators based on interpretable machine learning for the study areas still needed to be accomplished. For diverse regions, the reasons causing land subsidence may not be the same for each place at a single scale (Chitsazan et al., 2020), and a comprehensive set of scientific optimization techniques is needed to screen the features in specific study areas. The screening can make the training data more appropriate to the research areas and make the elements driving land subsidence more explanatory. The model prediction findings could be more scientific, improving prediction results.

To tackle these concerns, we implemented the SHAP values suggested by Lundberg and Lee (2017) to evaluate the key elements contributing to land subsidence in the specified region. SHAP (Shapley Additive exPlanations) is a technique that elucidates the predictions made by machine learning models. It relies on the cooperative game theory principles of the Shapley value notion. SHAP has been utilized across various domains in previous studies. Bhandari et al. (2020) developed a machine-learning model for intrusion detection systems (IDS) in communication. They used SHAP values to reduce computational complexity and maintain prediction performance. The algorithm selected the top 15 key features based on an index. Alenezi et al. (2021) utilized SHAP values to identify cyber security threats and explained the results of the machine learning model. SHAP values quantify the specific impact that each unique feature has on the model's prediction results. Despite their shortcomings, SHAP values remain a potent and extensively employed way of interpretation. They offer crucial insights that aid in comprehending the predictions made by machine learning models.

This study involved the construction of a BP neural network and random forest models to analyze the cumulative land subsidence in Chongqing and Chengdu from January 2022 to February 2023. The training results of these models were interpreted using SHAP values. Features with a stronger correlation to land subsidence were identified and used to retrain the model. Moreover, K-fold cross-validation was employed to validate the trained model and assess whether the model's predictive performance was enhanced, demonstrating the scientific and practical nature of our feature optimization and screening process. This study aims to investigate the following two inquiries: The objective of this study is twofold: (1) to suggest a series of optimization methods for identifying small-scale land subsidence characteristics that can be applied to various cities, and (2) to assess if the performance of the land subsidence prediction model, which is based on neural networks, is enhanced after the optimization of these features.

2 Study Area

2.1 Study Area

The study area of this research is the cities of Chengdu and Chongqing, two of the most representative cities in southwest China and one of the core cities in the Chengdu-Chongqing Economic Circle (China government network, 2016). Chengdu is located in the central part of the Sichuan Basin, China, between 103°36′ and 104° 53′ E longitude and 30°5′ and 31°26′ N latitude. Chongqing is located in the middle and lower reaches of the Yangtze River in China, between 105°17′ and 110°11′ E longitude and 28°10′ and 32°13′ N latitude, with a distance of about 300 km between the two cities. Both Chengdu and Chongqing belong to the humid subtropical climate zone, but due to their geographical location and topography, they have some differences in climatic conditions.

Chengdu is an inland plain city with relatively gentle terrain and surrounded by mountains. The soil in the Chengdu plain is fertile and suitable for agricultural development. Chengdu's climate is mild and humid, with four distinct seasons. Summers are short and rainy, with high average temperatures; winters are cold and dry, with low temperatures. On the other hand, Chongqing is one of the few mountainous cities with complex and diverse terrain. Its undulating terrain contains a large number of mountain ranges and hilly areas. Chongqing's mountainous terrain gives the city a unique landscape and natural environment and adds to urban planning and infrastructure development challenges. Chongqing's climate is hot and humid, with long, hot summers and relatively short, wet, and cold winters. Rainfall is abundant, with high average annual precipitation.

These two cities contain the vast majority of urban topography, with Chengdu known for its plain landscape and rich agricultural resources and Chongqing attracting visitors with its mountainous beauty and features. By studying the geographical characteristics of these two cities and the analysis of land subsidence data, relevant geological and urban development issues can be better understood and addressed. The study area selected for this research is the Chongqing and Chengdu metropolitan areas, and continuous monitoring and prediction are necessary for urban areas as areas with high incidences of land subsidence.

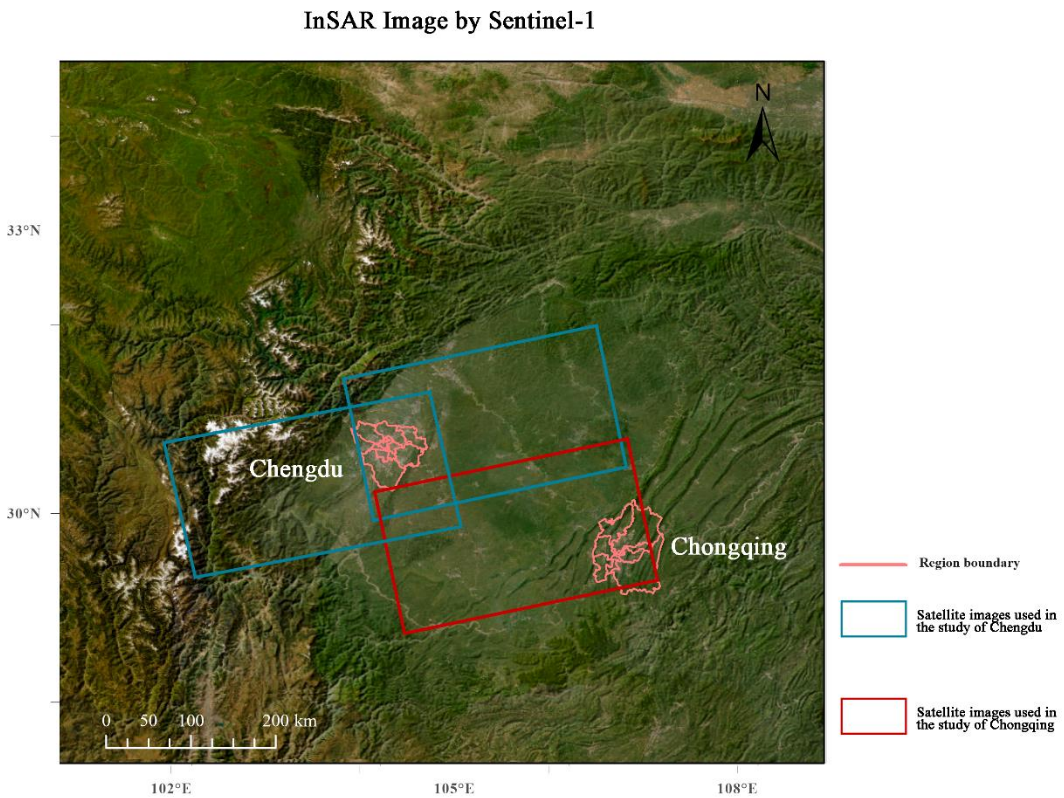

Figure 1. Selected InSAR Image by Sentinel-1

2.2 Datasets

This study uses the Sentinel-1 radar image set, which contains three ascending radar images acquired in January 2022 and three ascending radar images acquired in January 2023, to obtain ground subsidence monitoring results. This data is used to count where subsidence occurs between 2022 and 2023. Specific parameters are shown in Table 1.

Table 1. Parameters of Sentinel-1 data

Data Source	Sentinel-1
Polarization	VV
Orbit Direction	Ascending
Product Level	SLC
Resolution	30m
Time	January 7, 2022 and January 7, 2023

The main factors affecting subsidence include geological formations in the study area, urban construction (large commercial areas, high-rise buildings, subways), faults, hydrogeology, groundwater information, and rainfall. Therefore, we collected average temperature raster data, groundwater content data, surface evaporation data, and precipitation data obtained through the ERA5-Land dataset published by the European Union and the European Centre for Medium-Range Weather Forecasts Organization. We obtained the average data for 2022-2023 by raster calculation with ArcGIS. NDVI data is from MOD13A3 under NASA's MODIS dataset, and we obtained the month-by-month NDVI data from January 2022 to January 2023 and calculated the average NDVI data for 2022-2023. The metro data of major cities in China were produced by ourselves for the latest completed metro data in 2023. 2020 Chongqing and Chengdu metropolitan area-wide extensive building block data were provided by the National Tibetan Plateau Data Center (TPDC) Vectorized rooftop area data for 90 cities in China dataset. The 2023 water system data and 2023 road data were provided by Open Street Map (OSM), and our data query produced the fault zone data. The specific data attributes will be shown in Table 2. All of these data may be one of the factors that induce the occurrence of land subsidence, and we included all of them in the training data in order to cover a broader range of possible subsidence training features and to capture the most influential factors of land subsidence from a wide range of features.

Table 2. Attribute of data

Data name	Data Resolution	Data Phase	Data Source
Temperature Raster Groundwater Content Raster Surface Evaporation Raster Precipitation	9km	01/2022-01/2023	ERA5-Land dataset (Munoz Sabater, 2019)
NDVI	1km	1/1/2022-1/31/2023	MOD13A3 Dataset
Metro Data Fault Zone	-	2023	RESDC https://www.resdc.cn/

Large Building Block data	-	2020	National Tibetan Plateau Data Center (2021)
Water System Data	-	-	Open Street Map
Road Data	-	-	
Distance to Building	-	-	Data Processed by GIS software
Distance to Metro			
Distance to Road			
Distance to Water			
Distance to Fault Zone			

3 Method

The research framework is illustrated in Figure 2. Firstly, we processed the Sentinel-1 InSAR data using SNAP to obtain land subsidence data for Chengdu and Chongqing in 2022-2023. We divided the data into two categories to facilitate model training. The first category is the area where cumulative subsidence occurs greater than 0, and the second is where no subsidence occurs. Since the 30m resolution data is too large and cannot effectively express the information of subsidence occurring around it, we resampled the settlement data to 180m resolution to make the settlement data more generalizable. At the same time, we selected the distance to large buildings, distance to fault zones, distance to river systems, distance to subway lines, vegetation cover, rainfall, surface evaporation, underground runoff water content, temperature, and distance to roads as factors that may be related to the occurrence of land subsidence.

A total of 40,000 records (10,000 for Chongqing and 30,000 for Chengdu) were extracted from each record with and without subsidence.

After the data set is constructed, the model is trained by building a BP neural network and using K-Folds cross-validated to validate the model performance. The model is interpreted using SHAP to obtain the importance ranking of factors related to land subsidence. Then, we will construct the model validation set by taking 200 records each from Chongqing and Chengdu with and without subsidence in addition to these 40,000 records, for a total of 800 records. The model will be retrained from the top 3 to 7 most relevant features to evaluate the impact of different feature combinations on the model accuracy (Piles et al., 2021).

Finally, the above process obtained the optimal feature combination in the study data. Then, we compared the land subsidence prediction performance using the neural network with the random forest to observe whether the improvement occurred. The model we trained using this feature combination was predicted for the areas where subsidence did not occur in Chengdu and Chongqing, and the land subsidence possibility distribution map was obtained.

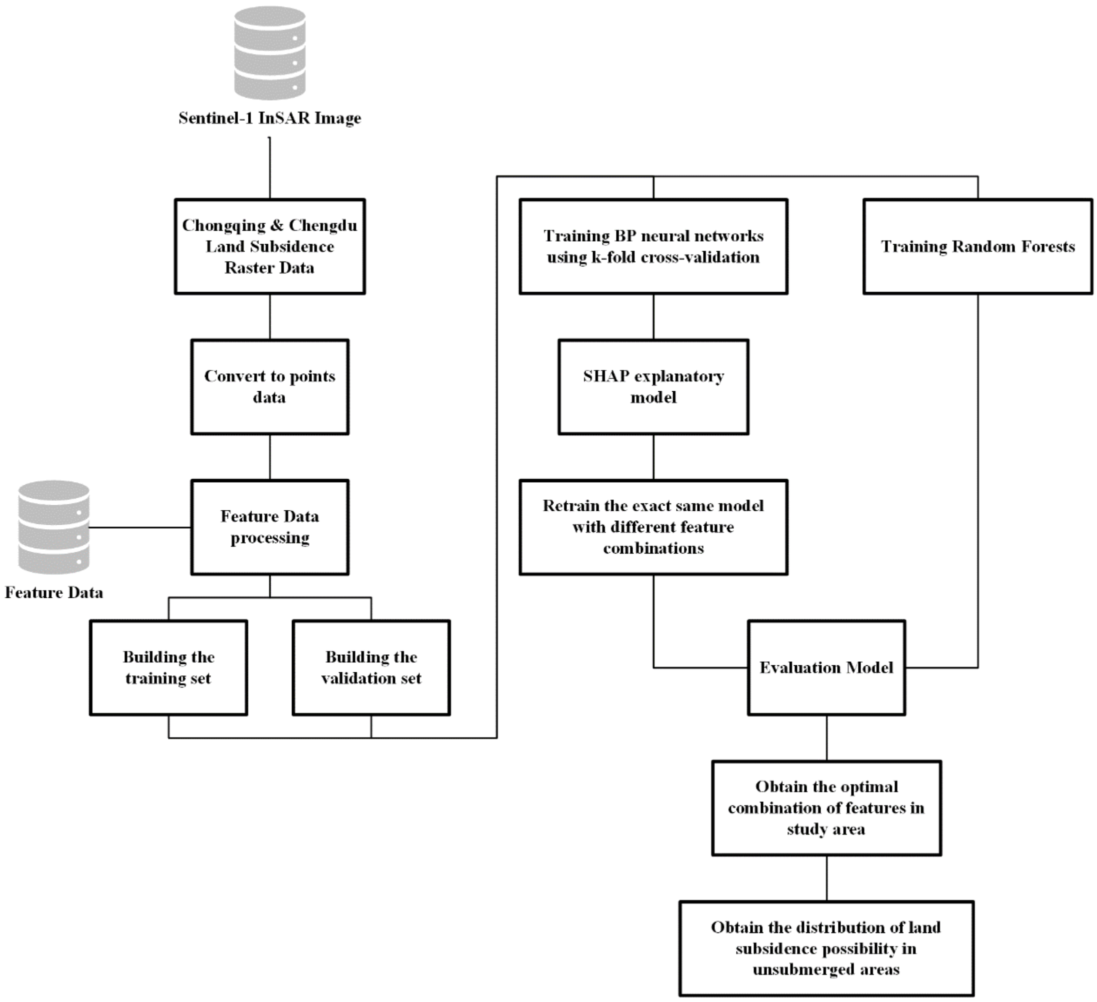

Figure 2. Feature Optimization Process

3.1 Land Subsidence Model

3.1.1 Random Forest

Random Forest (Breiman, 2001) is an integrated learning method that performs classification and regression tasks by combining multiple decision trees. Its core idea is to construct multiple decision trees by randomly selecting samples and feature subsets and combining their predictions to reduce the variance of the model and improve the generalization ability. The model can handle unbalanced data (Mingyue et al. 2021), balance the error by itself, and is insensitive to missing values and outliers, but overfitting can occur when the data is noisy.

3.1.2 BP Neural Network

BPNN (Backpropagation Neural Network) is a common artificial neural network model for solving classification and regression problem (Ang et al., 2022). It consists of an input layer, a hidden layer and an output layer, and is trained and weighted by a backpropagation algorithm. We built a BP neural network containing two hidden layers, using Adam (Wang et al., 2022) as the optimizer, ReLU as the hidden layer activation function, batch size of 64, and adaptive learning rate (liduka et al., 2021). The structure of the neural network is shown in figure 3.

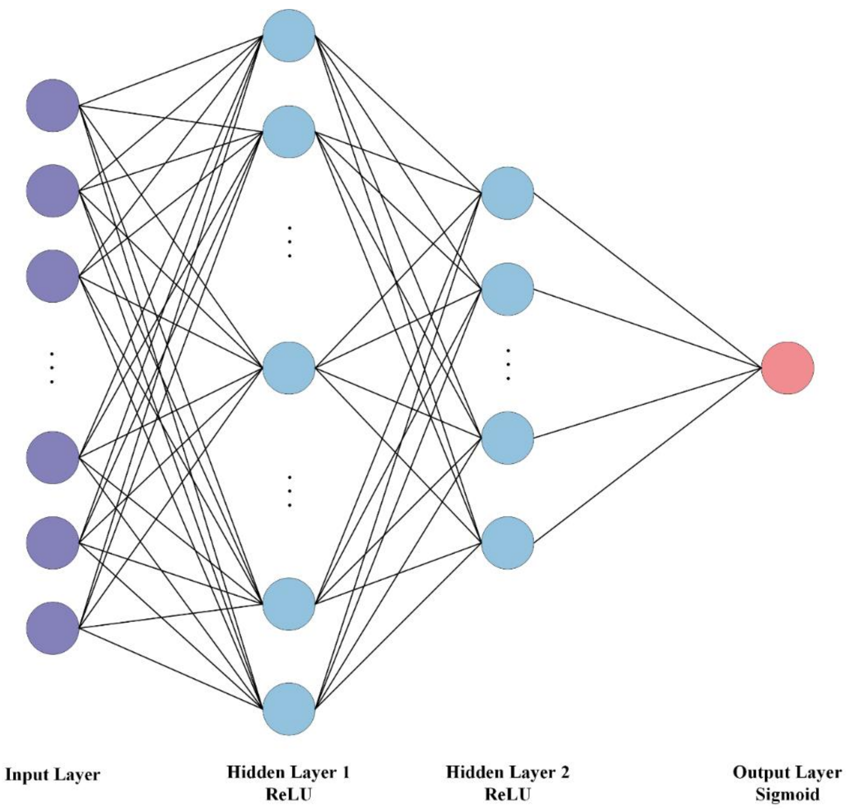

Figure 3. Structure of BPNN

3.2 Shapley Additive Explanations

SHAP (Shapley Additive Explanations) is a method for interpreting machine learning model predictions. It is based on the concept of Shapley values from cooperative game theory (Wang et al., 2022), and assigns a value to each feature to indicate the extent to which it contributes to the predicted outcome. SHAP values provide a more comprehensive, fair and consistent assessment of feature importance. Its fairness, interpretability, and feature interaction interpretation can provide a comprehensive scientific explanation of the relevance of land subsidence features in our model. The Shapley value is calculated as follows (Lundberg et al., 2017):

$$\phi_i(x) = \sum_{S \subseteq 1,2,\ldots,p \backslash i} \frac{|S|!(p-|S|-1)!}{p!}[f_x(S \cup i) - f_x(S)] \tag{1}$$

where *i* denotes the index of the feature to be interpreted, *x* is the input sample, and *p* is the total number of features. *S* is a subset of the features, $S \subseteq 1,2,\ldots,p \backslash i$ denotes the subset of features that does not contain feature *i*. $f_x(S)$ denotes the predicted output of the model given the input *x* and the subset of features *S*. $f_x(S \cup i)$ denotes the predicted output of the model given the input *x* and the subset of features $S \cup i$.

The formula indicates that for each feature *i*, iterates through all the subsets of features S that do not contain the feature and calculates the difference in the predicted output of the model between the two cases. By summing over all possible subsets and adjusting the weights according to the subset size, the contribution of feature *i* to the predicted outcome can be obtained. The SHAP value provides a fair and accurate way to account for the contribution of each feature, taking into account the interaction effect between features. A more comprehensive assessment of feature importance can be obtained by performing the calculation on multiple samples.

3.3 K-Fold cross-validation

K-fold cross-validation is a commonly used model evaluation method to assess the performance of machine learning algorithms and models (Wong et al., 2019). It divides the dataset into K equal-sized subsets, one of which is used as the test set and the remaining K-1 subsets are used as the training set. This process is repeated K times, each time selecting a different subset as the test set and the rest as the training set. The use of K-fold cross-validation allows for a better assessment of the stability and generalization ability of the model, providing more accurate results for model performance evaluation (Yang et al., 2021). We use a k-value of 5. The specific calculation steps are shown as figure 4:

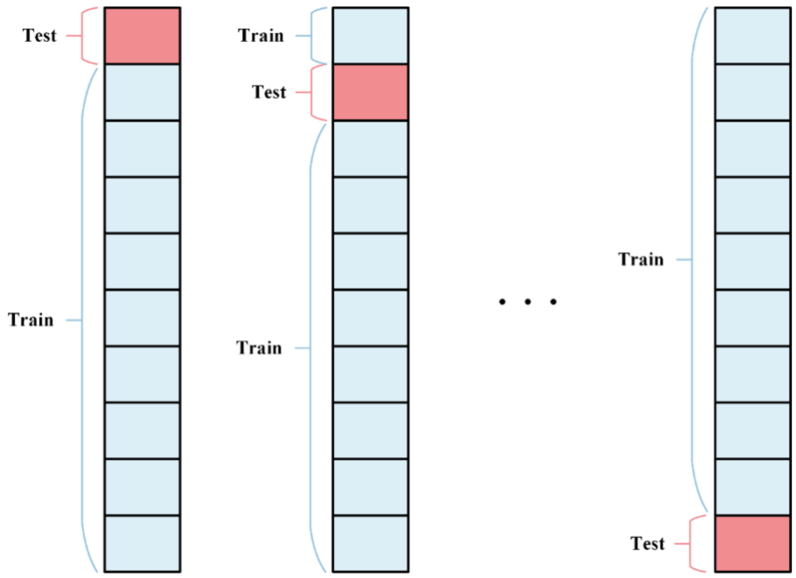

Figure 4. K-Fold cross-validation process

3.4 Model Evaluation Methodology

In this study, we use three metrics, MSE, MAE, and RMSE, to evaluate model prediction accuracy. MSE is one of the most commonly used assessment metrics in regression tasks. It calculates the square of the difference between the model's predicted values and the actual observed values and averages these differences. MSE imposes a greater penalty on larger values of prediction error, causing the model to focus more on those samples with larger prediction errors. The formula for MSE is as follows:

$$MSE = \frac{1}{n} \times \sum (y_i - \hat{y}_i)^2 \qquad (2)$$

where *n* is the number of samples, y_i is the actual observation, and $\hat{y}_i$ is the predicted value of the model. a smaller *MSE* indicates that the prediction of the model is closer to the actual observation.

MAE is another widely used metric for regression assessment. It calculates the absolute value of the difference between the model's predicted values and the actual observed values and averages these differences. The MAE is insensitive to outliers because it calculates the absolute value of the difference. This makes it more reliable in cases where outliers exist or where they have a large impact on the results. The formula for the MAE is given below:

$$MAE = \frac{1}{n} \times \sum | y_i - \hat{y}_i | \qquad (3)$$

RMSE is the square root of MSE. It measures the average difference between the model's predicted values and the actual observed values. RMSE combines the advantages of MSE, sensitivity to outliers, and intuitively meaningful results that match the units of the original data. The formula for RMSE is given below:

$$RMSE = \sqrt{MSE} \tag{4}$$

4 Results and Discussion

4.1 Land subsidence in the study area

In this study, we analyzed the land subsidence in Chengdu and Chongqing using InSAR data recorded by Sentinel-1. By looking at the 2022-2023 cumulative land subsidence data, in figure 5, we can observe the prominent locations where subsidence occurs. In Chengdu, land subsidence predominantly manifests in the southeastern sector of the city, notably in the Xizhiqui Mountain area (subsidence depth: 0.04m), the vicinity of Shuangliu International Airport (subsidence depth: 0.01m), Chengdu University Town (subsidence depth: 0.01m), and the eastern mountainous region (subsidence depth: 0.03m). For the Chongqing region, significant land subsidence is concentrated in the southeastern mountainous areas, the southern mountainous areas (0.068m), Guangyang Island (0.032m), the Yubei Industrial Park (0.026m) (which includes large factories such as Ford), and the Yubei Mining Park (0.035m).

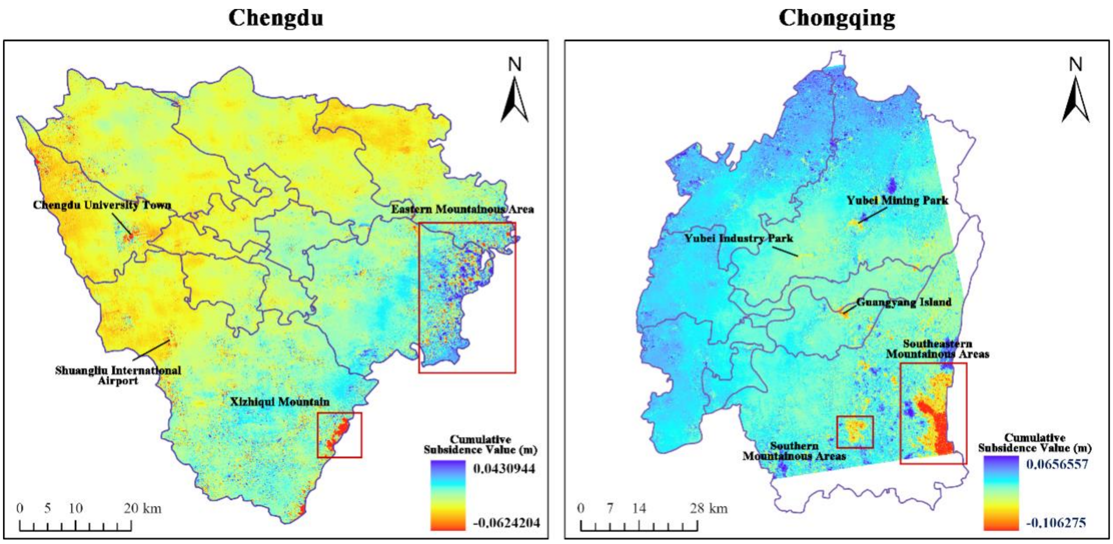

Figure 5. Actual land subsidence occurrence

4.2 Relationship between land subsidence and relative factors

4.2.1 Impact of river systems

Based on Figure 6, most subsidence occurrences are concentrated close to river systems, typically within a buffer distance ranging from approximately 3 to 6 km (Figure 6c). In Chengdu, the maximum subsidence near the river system, within a 1.5 km buffer zone, reaches 2.4 cm, encompassing an overall subsidence area of nearly 4.5 km^2—comprising almost 75% of the total buffer zone (Figure 6a). Figure 6b illustrates a growing severity in settlement issues, with a peak settlement value of 6.5 cm observed within a range of 2.9 kilometers.

Similarly, Chongqing exhibits pronounced subsidence near large lakes and rivers, particularly in the southwestern region of Guangyang Island (Figure 6c), where cumulative settlement reaches 6.5 cm, covering a subsidence area exceeding 100 km^2. Another notable subsidence area is near the Jialing River, with a land subsidence value of 5.9 cm and a total area of about 10 km². Generally, the groundwater resources around ponds tend to be relatively abundant, as ponds attract groundwater for recharge. Consequently, the elevation in groundwater levels may contribute to land subsidence (Rajabi et al., 2016).

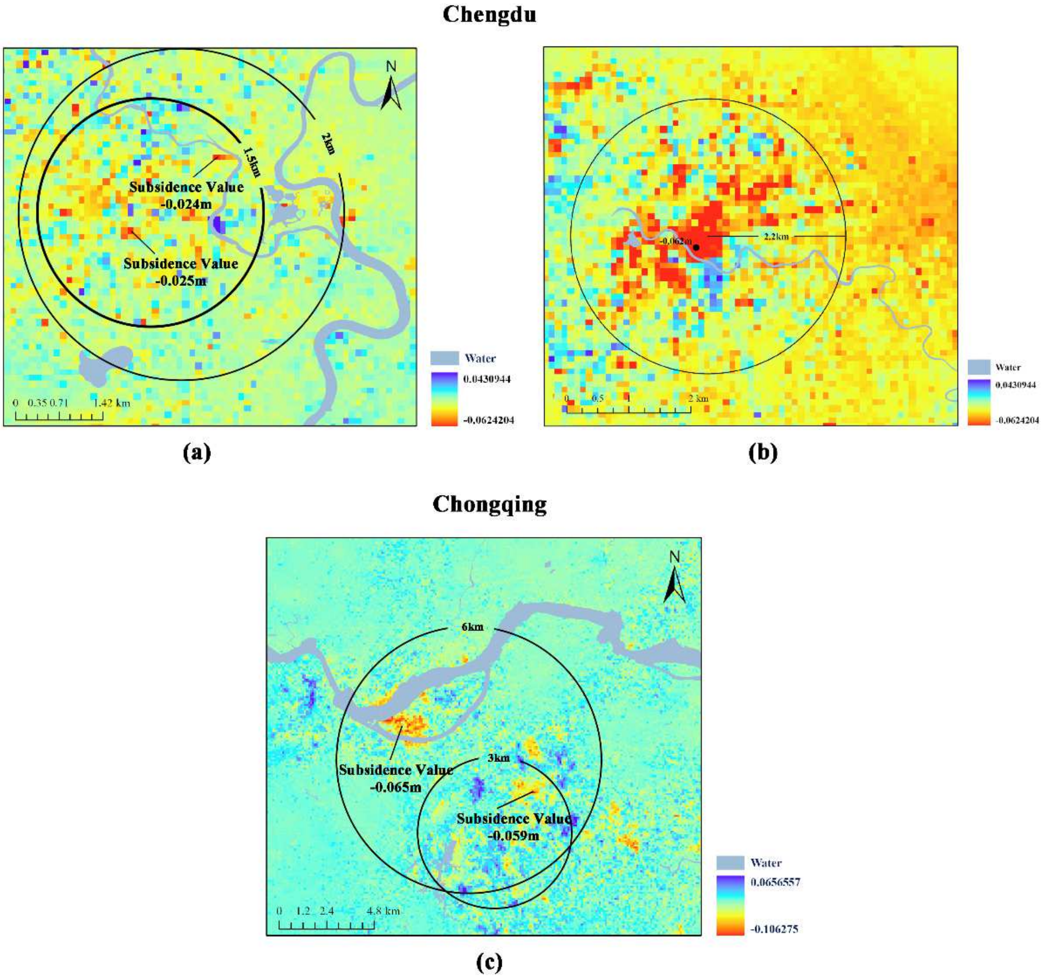

Figure 6. Links between land subsidence areas and river systems

4.2.3 Impact of large buildings

As illustrated in Figure 7, the regions experiencing land settlement in Chongqing and Chengdu are closely associated with large buildings, particularly within a 2km buffer distance. The maximum cumulative subsidence value observed in Chengdu is 4.8 cm, mirroring a comparable value in Chongqing at 5.7 cm. Notably, the cumulative subsidence encompasses an area of 6 km^2 near large buildings. Remarkably, settlements are exclusively discerned in areas featuring large buildings; however, their prevalence is notably heightened in zones surrounded by such structures.

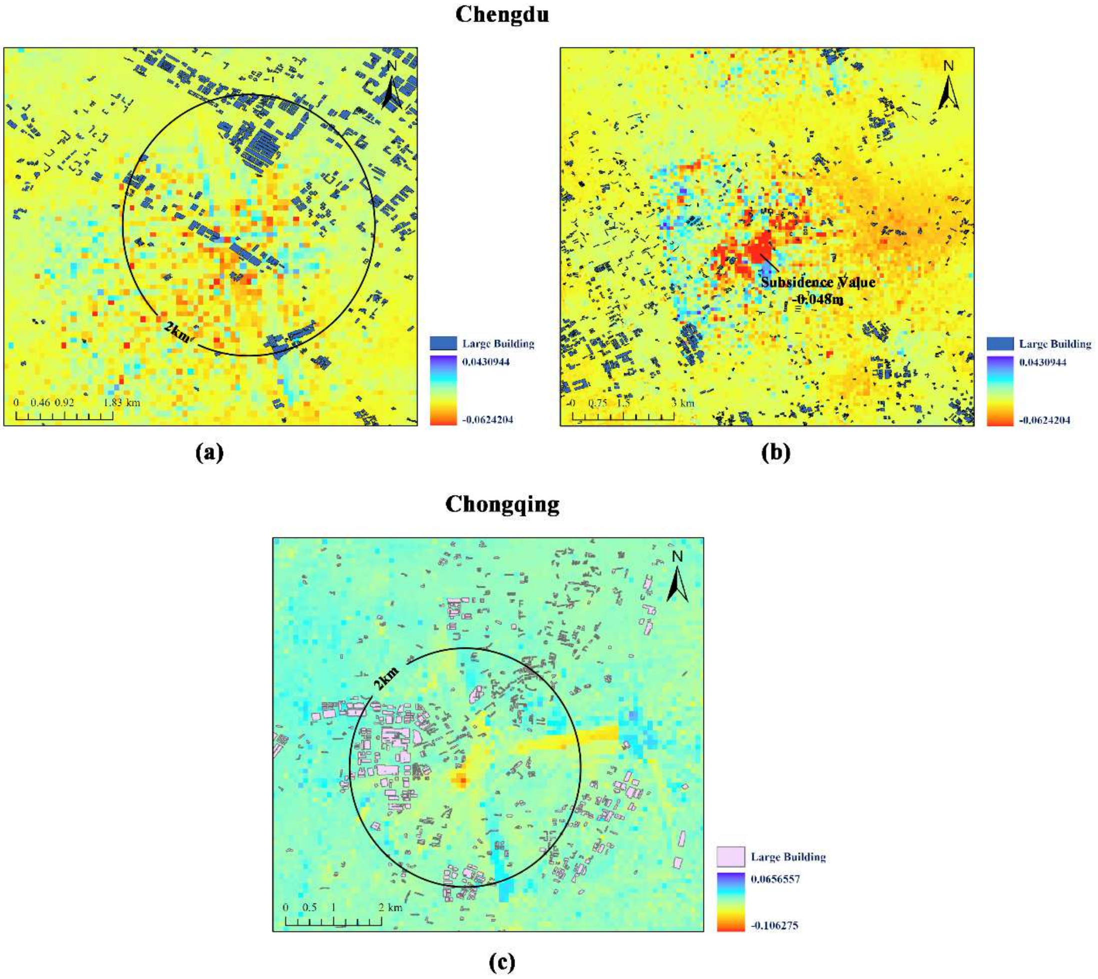

Figure 7. Links between land subsidence areas and large building

4.2.4 Impact of fault zones

Furthermore, as depicted in Figure 8, empirical evidence indicates a significant correlation between actual land subsidence and fault zone areas (Mohammaday et al., 2019). In Chongqing, the cumulative settlement value within the subsidence zone adjacent to the fault zone reached 0.048 m, with the closest proximity to the zone being 3.5 km. The mountainous terrains proximate to fault zones exhibit an elevated susceptibility to subsidence. This phenomenon is similarly observed in Chengdu (Figure 8a), where subsidence is manifest in mountainous areas, registering a depth of 0.018m at a distance of 2.7 km from the fault zone.

Given Chongqing's topography as a mountainous city, as figure 8b shows, it has undergone extensive subsidence in the fault zone region, covering an area of 425 km²—approximately 2.89 times larger than Chengdu's subsidence area of 147 km².

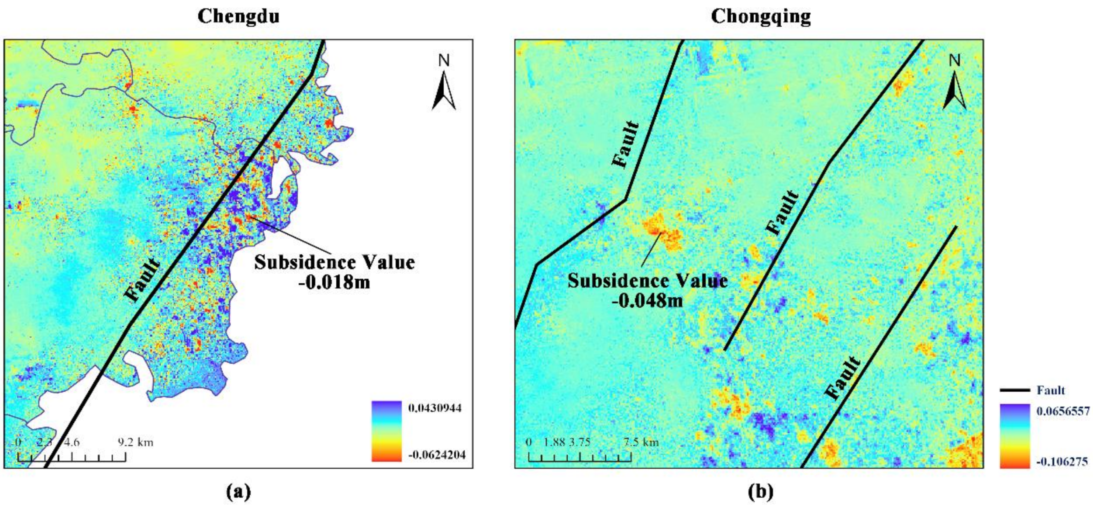

Figure 8. Links between land subsidence areas and fault zone

4.3 Land Subsidence Model Explanation

By inputting the previously constructed land subsidence dataset of Chengdu and Chongqing into the model for training. We applied the K-fold cross validation to train the neural network, and trained a total of five models. Finally, we took the average of the model evaluation metrics as the results of the evaluation of the model trained with the current features, and the model training results are as follows (figure 9).

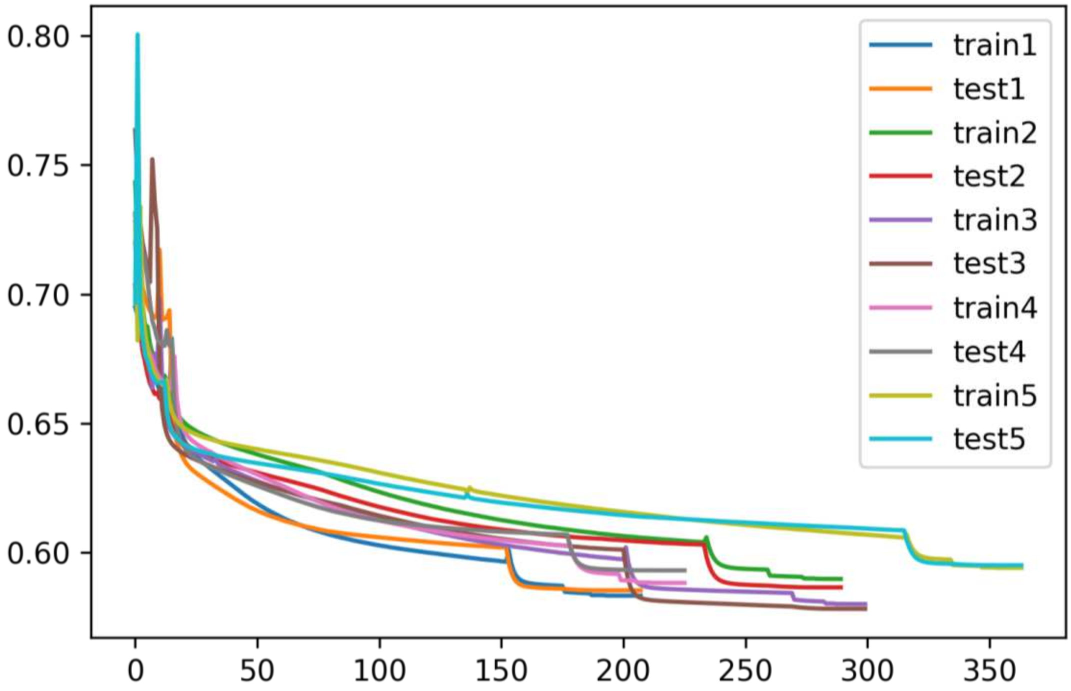

Figure 9. Training loss of BP Nerual Network

4.3.1 Feature Importance Analysis

After the SHAP interpreter shown as figure 10, the neural network training feature importance was generated. At the same time, we also input the same data into the random forest model.

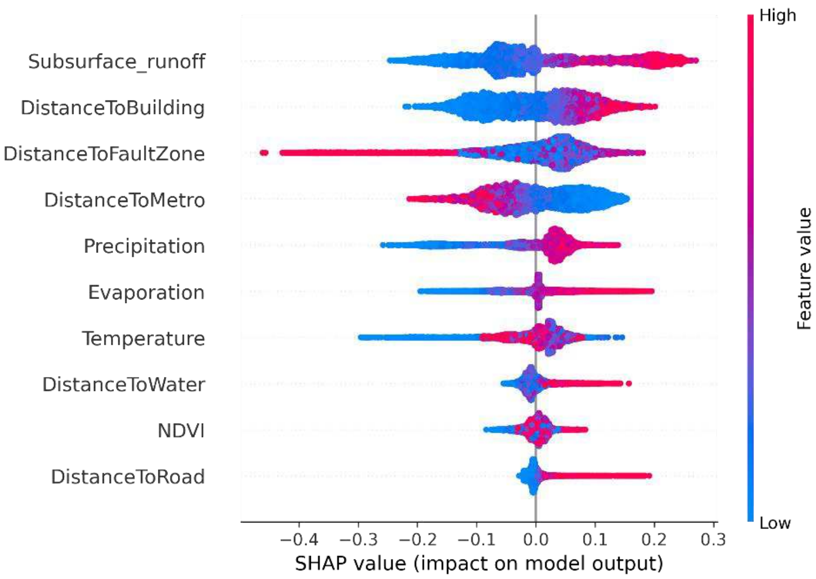

Figure 10. Features importances evaluate by SHAP value

Taken together, subsurface runoff has a significant impact on soil moisture and water supply, and therefore plays a key role in land subsidence prediction, consistent with the findings of some previous studies (Yu et al., 2019; Chen et al., 2022; Wang et al., 2019). Among them, the SHAP value for subsurface runoff reaches about 0.3. Whereas, the distance from large buildings in most of the areas exerts a positive influence on the model, with the SHAP value for the effect on the model concentrating in the range of 0-0.1. Also, It can be seen that the distance to fault zone has a large negative effect on the prediction results, the minium SHAP value reach -0.4, and the greater the distance from the fault zone, the greater the negative effect on the model results. But because there are fewer zones that have a higher negative effect, it does not have as large an effect on the model as the other two factors, even though it has the largest SHAP value. Their SHAP values are all between -0.3 and 0.2. The distance to the river system and the distance to the road are basically positively correlated with the model, i.e., the larger the distance, the stronger the positive effect on the model.

4.3 Feature optimization by SHAP value

We randomly selected 800 land subsidence data outside the training set, and randomly selected an equal number of 200 data in Chengdu and 200 data in Chongqing with and without subsidence as the validation set. The AVG MSE of the model without previous feature screening using K-fold cross-validation is 0.217, the AVG MAE is 0.438, and the AVG RMSE is 0.466.

Table 3. Evaluation of the effectiveness of different models

Model	MAE	MSE	RMSE
Random Forest	0.31	0.31	0.557
BP Neural Network (Includes all features)	0.219	0.219	0.468
BP Neural Network (Includes top 3 features)	0.197	0.197	0.444
BP Neural Network (Includes top 4 features)	**0.190**	**0.19**	**0.436**
BP Neural Network (Includes top 5 features)	0.216	0.215	0.464

BP Neural Network (Includes top 6 features)	0.216	0.216	0.465
BP Neural Network (Includes top 7 features)	0.229	0.228	0.478

Meanwhile, we selected the top 3, 4, 5, 6, and 7 features in the importance ranking of features generated by SHAP, respectively, and re-trained the exact same model as before using K-fold cross validation (with the same model structure and model weights), and evaluated the prediction effect of the model using the metrics such as average MSE, MAE, and RMSE generated by K-fold cross validation. The first seven features are: subsurface_runoff, DistToBuilding, DistToFaultZone, DistToMetro, Precipitation, Evaporation, and Temperature. the specific model evaluation results are shown in Table 3.

Through feature screening, we found that selecting the first 4 features to train the model model prediction performance is the best, and its average MSE is improved by 38.7% compared to the random forest, and 13.2% compared to the neural network model without feature screening. The average RMSE values were improved by 36.8% and 6.8%, respectively. The average MAE value is improved by 5.9% compared to the model without feature screening.

4.5 Prediction of potential subsidence area

We have identified non-subsided areas from the land subsidence datasets of Chengdu and Chongqing. Subsequently, we utilized a pre-trained neural network model, which had undergone feature screening, to predict the probability of potential subsidence in these non-subsided regions.

4.5.1 Analyze the results of potential subsidence prediction

Using the model trained after feature screening, we obtain the probability of subsidence in the non-subsidence areas of Chongqing and Chengdu. For Chengdu, the model predicts that the areas with higher probability of land subsidence are mainly in the western industrial area and the eastern mountainous area, while the probability of subsidence in the main urban area is lower, which is consistent with the actual situation.

In the western industrial area, there are a larger number of large factories and plants, and large buildings usually have high self-weight, by exerting pressure on the foundation. Over a long period of time, the weight of the building can cause the soil to settle gradually. Especially if the buildings have heavy structures or are constructed on weak soil foundations, their impact on the soil is more significant and may lead to settlement. Also, groundwater pumping may be more frequent in industrial areas with large buildings. When groundwater is overpumped, the water table drops, resulting in less pore water in the soil, which increases the risk of soil consolidation and settlement. The model predicts a higher probability of possible subsidence in this area.

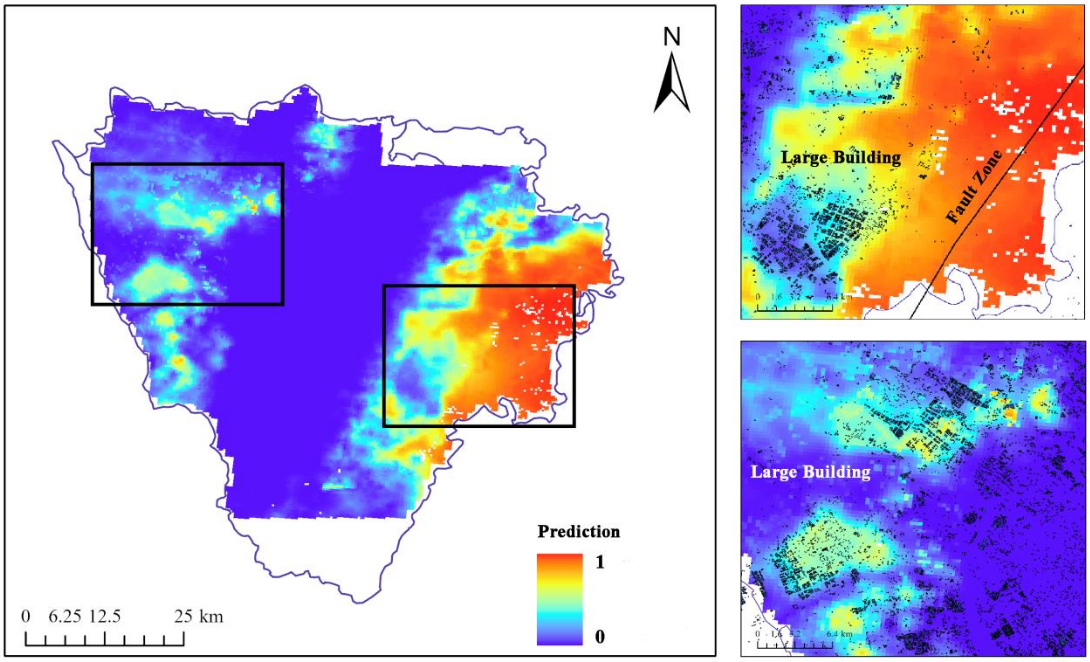

Figure 11. Predicted areas of possible subsidence in Chengdu

In the eastern mountainous areas, located in the fault zone region, the model predicts a very high probability of possible subsidence (figure 11). Geologic formations in mountainous areas are generally complex, with a variety of rock formations, faults and folds, and other geologic structures. These geologic formations are usually non-uniform and may undergo stress release or deformation on time scales. Due to the instability of the geologic formations, soils and rocks in mountainous areas are susceptible to displacement and deformation, resulting in subsidence (Doke et al., 2020). There are also a large number of large buildings in this area, and the model does not consider that there is a greater likelihood of land subsidence in this area, indicating that the model is not focusing on these features of large buildings or fault zones, but rather a comprehensive consideration with good generalization performance.

By modeling the predicted areas of possible subsidence in Chongqing shown as figure 15, we can observe that the main areas of subsidence occur in urban areas, which is consistent with the actual situation. As a mountainous city, Chongqing has a complex geological structure in the urban area, which is more prone to subsidence.

According to the analysis and prediction results, the factors that may induce land subsidence in Chongqing are fault zones and river systems. However, it does not mean that the model only determines the probability of land subsidence based on these two factors. We can observe that in the western suburbs of Chongqing, near the fault zone area, the model also does not consider the probability of subsidence to be high, which is a good illustration of the generalization of the model, which can analyze the probability of land subsidence based on the actual situation of the urban area.

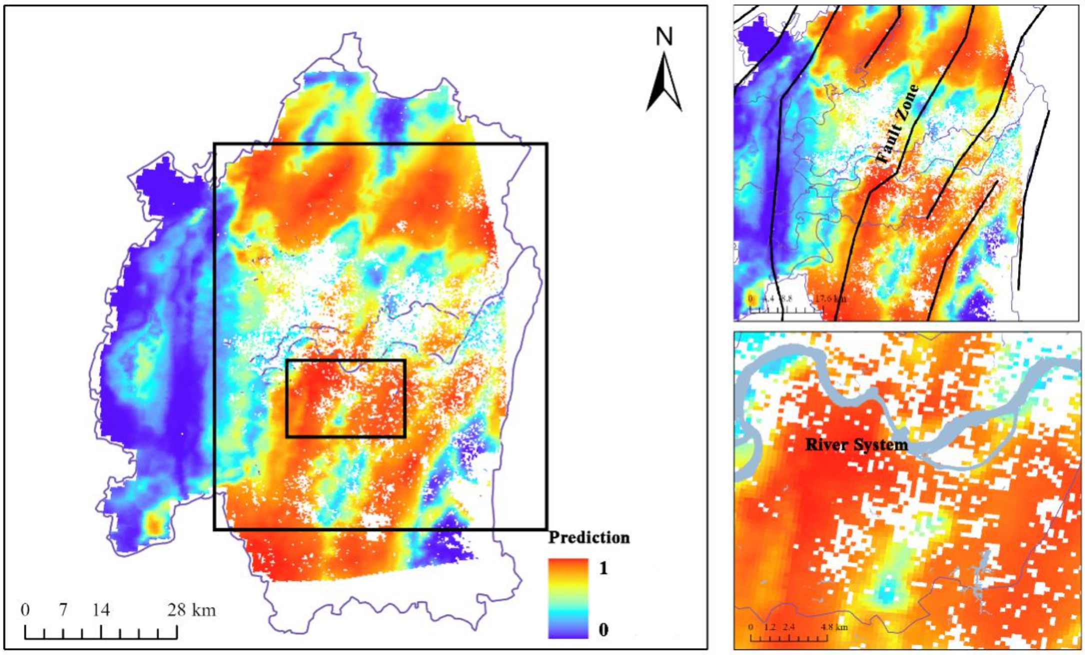

Figure 12 Predicted areas of possible subsidence in Chongqing

4.5.2 Analyze the factors induce the occurrence of potential subsidence

In order to explore the relationship between different features in the model directly and the prediction results, and to better understand the basis of the model prediction, we plotted the correlation between the features used in the training model and the prediction results, from the relationship between the subsurface runoff and the probability of land subsidence as shown in Figure 13. From that figure, we can observe that the probability of possible subsidence in Chengdu is in the area with a groundwater content of 0.001-0.0015 m^3 maximum, and the subsidence area is mainly concentrated in the area with groundwater content of 0.0005-0.00015. A different result is observed in Chongqing, where the main subsidence areas are those with groundwater contents between 0.00075 m^3 and 0.0008 m^3.

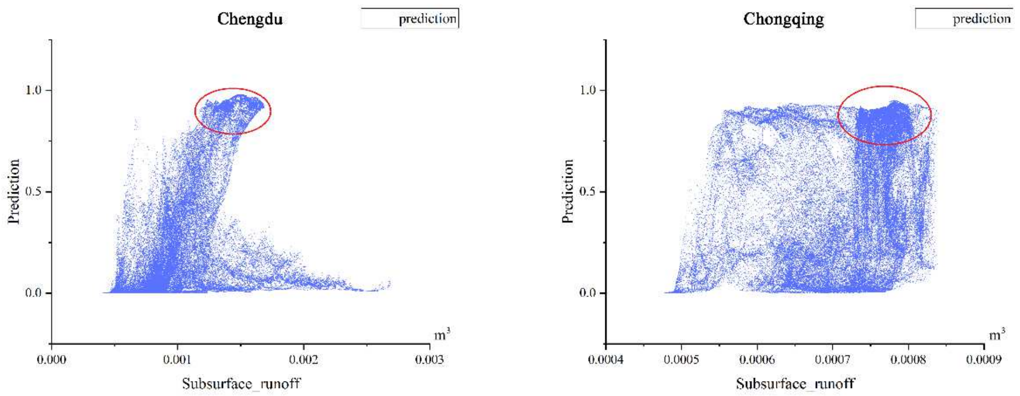

Figure 13. Relationship between prediction and subsurface runoff

We posit that the observed phenomenon may be attributed to an increase in groundwater content, which can lead to soil reaching a saturated state, where all pores are filled with water. In such a saturated state, soil mobility and compressibility are relatively reduced, consequently mitigating the degree of settlement. Additionally, the elevation in groundwater content results in an increase in pore water pressure within the soil, which, in turn, applies effective stresses on the soil body, thereby diminishing soil consolidation and settlement 3.

Besides sub-surface runoff, we also explored the relationship between subsidence probability and the proximity distance to the fault zones. The figure 14 reveals a concentration of subsidence areas in Chengdu where settlement is likely within a 15 km radius from the large building. A distinct demarcation line emerges, with a majority of high probability settlement areas situated nearer to the fault zone. In the case of Chongqing, which is a mountainous city with most urban areas close to the fault zones, the result is not as obvious as Chengdu but we can still tell that high probability of subsidence areas are relatively mainly clustered within a 6 km distance inside the fault zones. This finding aligns with the empirical analysis, underscoring that areas closer to the fault zone exhibit an increased susceptibility to land subsidence.

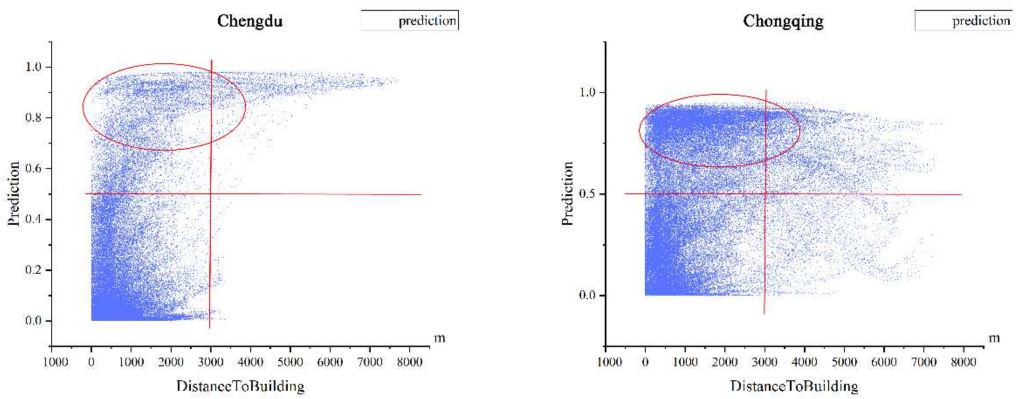

Figure 14. Relationship between prediction and distance to building

Subsequently, we conducted an analysis on the relationship between model-predicted subsidence areas and their proximity to large buildings (figure 15). Unlike other features, the association between subsidence probability and distance to large buildings is not overtly apparent. However, in both cities, areas exhibiting a high potential subsidence are mostly distributed within a 3km radius to large buildings. Similarly, the model displays limited sensitivity to subway distance (figure 16), and the prediction results with relatively high values (⩾0.5) are generally distributed with very limited distinct clustering patterns. Our analysis suggests that the lack of sensitivity may be attributed to these two features potentially serving as auxiliary factors, enhancing the model's generalization rather than exerting a direct influence.

In the development of prediction models, auxiliary features are occasionally introduced. While they may lack a strong direct correlation with target variables, they contribute supplementary information through interaction with other features. These auxiliary features enable the model to capture intricate patterns and correlations within the data, augmenting its overall generalization capability. Despite the absence of a direct correlation with settlement probability, these features may positively impact the model's predictive capacity when integrated with other relevant features.

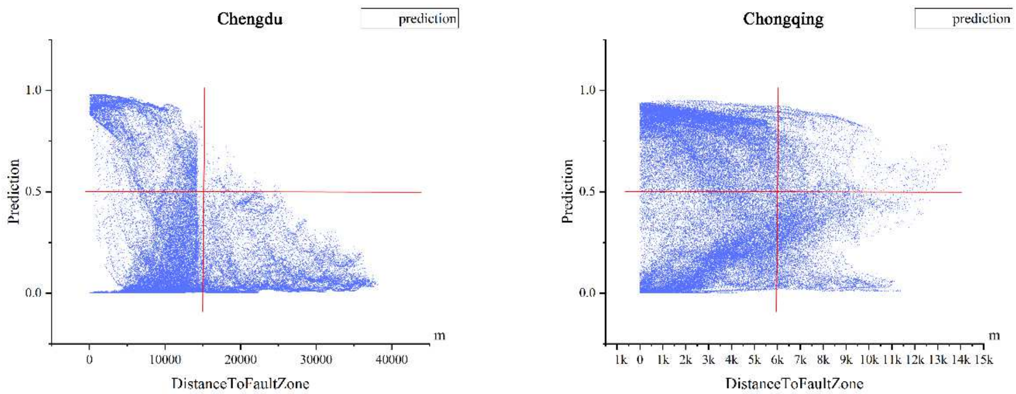

Figure 15. Relationship between prediction and distance to fault zone

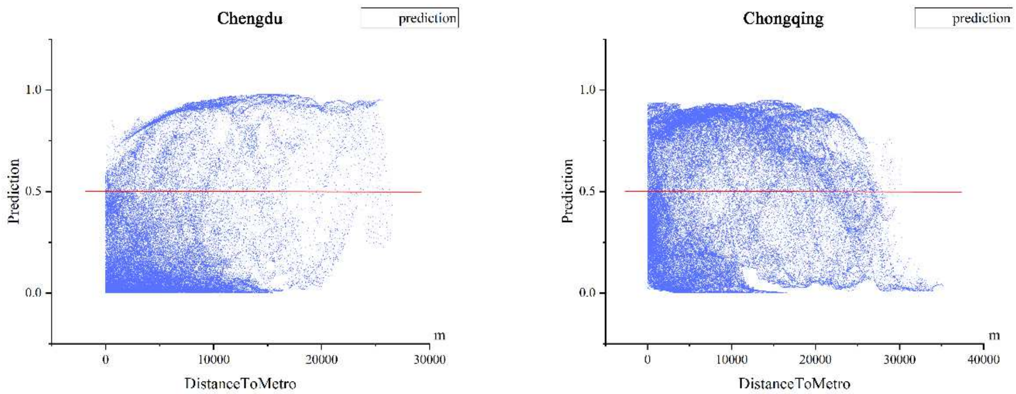

Figure 14. Relationship between prediction and distance to metro

5 Conclusion

None of the current research on land subsidence prediction has a rigorous and standardized process for optimizing the selection of training features to generate the most accurate model predictions. The inappropriate selection of training samples directly impacts the accuracy of the final results in predicting large-scale urban land subsidence. Conducting feature optimization is an essential and crucial stage. Hence, to address the issue of optimizing features in predicting land subsidence, we utilized six Sentinel-1 InSAR images to acquire the cumulative subsidence dataset for Chengdu and Chongqing from 2022 to 2023. This study introduces a feature optimization technique that utilizes SHAP values and neural networks. The method was applied to predict land subsidence in Chengdu and Chongqing. The results demonstrate that the optimized model outperforms both the traditional tree model and the neural network model without feature screening. Specifically, the optimized model's average mean squared error (MSE) is improved by 38.7% compared to the Random Forest model and by 13.2% compared to the neural network model without feature screening. Model that does not involve the process of feature screening.

Simultaneously, an analysis was conducted on the factors that could cause land subsidence. The findings revealed that the level of groundwater content is the most significant element influencing land subsidence. Furthermore, the four most crucial attributes, namely subsurface runoff, DistToBuilding, DistToFaultZone, and DistToMetro, were chosen to train the land subsidence prediction model to achieve optimal prediction outcomes at this research scale.

Upon analyzing the model's prediction results for the regions of Chengdu and Chongqing susceptible to subsidence, we observed that the model yielded varying predictions for different areas. These predictions were derived based on the actual conditions of each respective region. It demonstrates that the model exhibits strong generalization and avoids issues such as overfitting by reducing the amount of characteristics.

We utilized Sentinel-1 InSAR data to analyze Chongqing and Chengdu's recorded land subsidence data from 2022 to 2023. Through this analysis, we developed an optimized process for predicting the characteristics of the land subsidence model. As a result, we successfully identified the potential areas where land subsidence is likely to occur in the future. This is a novel concept for future land subsidence prediction model investigation. Through empirical validation, we have demonstrated that this technique can efficiently identify traits and provide comprehensive explanations for the selection of settlement positions. Following the feature selection process, the model's predictive performance is enhanced while the computational workload significantly decreases. Furthermore, the neural network exhibits superior predictive accuracy in feature selection compared to the random forest model. By establishing a solid theoretical foundation, we may effectively help the prevention of land subsidence in the future.

Funding Declaration： This work was supported by the Natural Science Foundation of Chongqing (Grant numbers Nos.cstc2020jcyj-msxmX0517).

Data availability statement: Raw data were generated at NASA EARTH OBSERVATION DATA. Derived data supporting the findings of this study are available from the corresponding author Xiancai Xiong on request.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

Reference

Larson, K. J., Başağaoğlu, H., & Marino, M. A. (2001). Prediction of optimal safe ground water yield and land subsidence in the Los Banos-Kettleman City area, California, using a calibrated numerical simulation model. *Journal of hydrology*, *242*(1-2), 79-102.

Kumar, S., Kumar, D., Donta, P. K., & Amgoth, T. (2022). Land subsidence prediction using recurrent neural networks. *Stochastic Environmental Research and Risk Assessment*, *36*(2), 373-388.

Wu, Y., You, X., & Zha, S. (2020). Mechanical behavior analysis of buried polyethylene pipe under land subsidence. *Engineering Failure Analysis*, *108*, 104351.

Yu, Q., Wang, C., Dai, Z., Ran, X., Amooie, M. A., Soltanian, M. R., & Soltanian, M. (2020). Effects of surface loading on groundwater flow and skeletal deformation. *Water Supply*, *20*(1), 287-295.

Gu, Z. Q., Gong, H. L., Zhang, Y. Q., Lu, X. H., Wang, S., Wang, R., & Liu, H. H. (2014). Research on monitoring land subsidence in Beijing plain area using PS-INSAR technology. *Spectroscopy and Spectral Analysis*, *34*(7), 1898-1902.

He, X. C., Yang, T. L., Shen, S. L., Xu, Y. S., & Arulrajah, A. (2019). Land subsidence control zone and policy for the environmental protection of Shanghai. *International journal of environmental research and public health*, *16*(15), 2729.

Ohenhen, L. O., & Shirzaei, M. (2022). Land subsidence hazard and building collapse risk in the coastal city of Lagos, West Africa. *Earth's Future*, *10*(12), e2022EF003219.

Park, J. H., Kang, J., Kang, J., & Mun, D. (2022). Machine-learning-based ground sink susceptibility evaluation using underground pipeline data in Korean urban area. *Scientific Reports*, *12*(1), 20911.

Saber, M., Abdel-Fattah, M., Kantoush, S. A., & Sumi, T. (2018). Implications of land subsidence due to groundwater over-pumping: monitoring methodology using GRACE data. *GEOMATE Journal*, *14*(41), 52-59.

Zou, L., Kent, J., Lam, N. S. N., Cai, H., Qiang, Y., & Li, K. (2015). Evaluating land subsidence rates and their implications for land loss in the lower Mississippi River basin. *Water*, *8*(1), 10.

Chen, M., Tomás, R., Li, Z., Motagh, M., Li, T., Hu, L., ... & Gong, X. (2016). Imaging land subsidence induced by groundwater extraction in Beijing (China) using satellite radar interferometry. *Remote Sensing*, *8*(6), 468.

Zhou, C., Gong, H., Chen, B., Gao, M., Cao, Q., Cao, J., ... & Shi, M. (2020). Land subsidence response to different land use types and water resource utilization in Beijing-Tianjin-Hebei, China. *Remote Sensing*, *12*(3), 457.

Guo, L., Gong, H., Ke, Y., Zhu, L., Li, X., Lyu, M., & Zhang, K. (2021). Mechanism of land subsidence mutation in Beijing plain under the background of urban expansion. *Remote Sensing*, *13*(16), 3086.

Li, F., Liu, G., Tao, Q., & Zhai, M. (2023). Land subsidence prediction model based on its influencing factors and machine learning methods. *Natural Hazards*, *116*(3), 3015-3041.

Ebrahimy, H., Feizizadeh, B., Salmani, S., & Azadi, H. (2020). A comparative study of land subsidence susceptibility mapping of Tasuj plane, Iran, using boosted regression tree, random forest and classification and regression tree methods. *Environmental Earth Sciences*, *79*, 1-12.

Wang, H., Jia, C., Ding, P., Feng, K., Yang, X., & Zhu, X. (2023). Analysis and prediction of regional land subsidence with InSAR technology and machine learning algorithm. *KSCE Journal of Civil Engineering*, *27*(2), 782-793.

Ma, P., Zhang, F., & Lin, H. (2020). Prediction of InSAR time-series deformation using deep convolutional neural networks. *Remote sensing letters*, *11*(2), 137-145.

Zhou, D., Zuo, X., & Zhao, Z. (2022). Constructing a large-scale urban land subsidence prediction method based on neural network algorithm from the perspective of multiple factors. *Remote Sensing*, *14*(8), 1803.

Shi, L., Gong, H., Chen, B., & Zhou, C. (2020). Land subsidence prediction induced by multiple factors using machine learning method. *Remote Sensing*, *12*(24), 4044.

Li, F., Liu, G., Gong, H., Chen, B., & Zhou, C. (2022). Assessing Land Subsidence-Inducing Factors in the Shandong Province, China, by Using PS-InSAR Measurements. *Remote Sensing*, *14*(12), 2875.

Rafiei Sardooi, E., Pourghasemi, H. R., Azareh, A., Soleimani Sardoo, F., & Clague, J. J. (2022). Comparison of statistical and machine learning approaches in land subsidence modelling. *Geocarto International*, *37*(21), 6165-6185.

Chitsazan, M., Rahmani, G., & Ghafoury, H. (2020). Investigation of subsidence phenomenon and impact of groundwater level drop on alluvial aquifer, case study: Damaneh-Daran plain in west of Isfahan province, Iran. Modeling Earth Systems and Environment, 6, 1145-1161.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Bhandari, S., Kukreja, A. K., Lazar, A., Sim, A., & Wu, K. (2020, June). Feature selection improves tree-based classification for wireless intrusion detection. In *Proceedings of the 3rd International Workshop on Systems and Network Telemetry and Analytics* (pp. 19-26).

Alenezi, R., & Ludwig, S. A. (2021, December). Explainability of cybersecurity threats data using SHAP. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 01-10). IEEE.

China government network (2016). Leading the new urbanization and agricultural modernization in western China through the Chengdu-Chongqing Urban Agglomeration Development Plan. https://www.gov.cn/guowuyuan/2016-03/30/content_5059846.htm (Accessed on 07-12-2023)

Muñoz Sabater, J. (2019): ERA5-Land monthly averaged data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.68d2bb30 (Accessed on 06-07-2023)

Nanjing, N. (2021). Vectorized rooftop area data for 90 cities in China (2020). National Tibetan Plateau Data Center. https://doi.org/10.11888/Geogra.tpdc.271702. https://cstr.cn/18406.11.Geogra.tpdc.271702. (Accessed on 06-07-2023)

Piles, M., Bergsma, R., Gianola, D., Gilbert, H., & Tusell, L. (2021). Feature selection stability and accuracy of prediction models for genomic prediction of residual feed intake in pigs using machine learning. Frontiers in genetics, 12, 611506.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Mingyue, F., Zao, F., Xiaodong, W., & Jun, M. (2021, May). A Pipeline Blockage Identification Model Learning from Unbalanced Datasets Based on Random Forest. In 2021 33rd Chinese Control and Decision Conference (CCDC) (pp. 696-701). IEEE.

Ang, K. M., Chow, C. E., El-Kenawy, E. S. M., Abdelhamid, A. A., Ibrahim, A., Karim, F. K., ... & Lim, W. H. (2022). A Modified Particle Swarm Optimization Algorithm for Optimizing Artificial Neural Network in Classification Tasks. Processes, 10(12), 2579.

Wang, Y., Xiao, Z., & Cao, G. (2022). A convolutional neural network method based on Adam optimizer with power-exponential learning rate for bearing fault diagnosis. Journal of Vibroengineering, 24(4), 666-678.

Iiduka, H. (2021). Appropriate learning rates of adaptive learning rate optimization algorithms for training deep neural networks. IEEE Transactions on Cybernetics, 52(12), 13250-13261.

Wang, Y., Lang, J., Zuo, J. Z., Dong, Y., Hu, Z., Xu, X., ... & Li, H. (2022). The radiomic-clinical model using the SHAP method for assessing the treatment response of whole-brain radiotherapy: a multicentric study. European Radiology, 32(12), 8737-8747.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Wong, T. T., & Yeh, P. Y. (2019). Reliable accuracy estimates from k-fold cross validation. IEEE Transactions on Knowledge and Data Engineering, 32(8), 1586-1594.

Yang, J., Wu, Z., Wang, Z., Zhang, D., Wang, W., Wen, Q., & Zhang, W. (2023). Enhanced anisotropic radius basis function metamodel based on recursive evolution Latin hypercube design and fast K-fold cross-validation. Structural and Multidisciplinary Optimization, 66(7), 169.

Missaoui, R., Abdelkarim, B., Ncibi, K., Gentilucci, M., Brahmi, S., Ayadi, Y., & Hamed, Y. (2023). Mapping groundwater recharge potential zones in arid region using remote sensing and GIS perspective, Central Tunisia. Euro-Mediterranean Journal for Environmental Integration, 1-15.

Yang, T., Ala, M., Guan, D., & Wang, A. (2021). The effects of groundwater depth on the soil evaporation in Horqin Sandy Land, China. Chinese Geographical Science, 31, 727-734.

Rajabi, A. M., & Ghorbani, E. (2016). Land subsidence due to groundwater withdrawal in Arak plain, Markazi province, Iran. Arabian Journal of Geosciences, 9, 1-7.

Mohammady, M., Pourghasemi, H. R., & Amiri, M. (2019). Land subsidence susceptibility assessment using random forest machine learning algorithm. Environmental Earth Sciences, 78, 1-12.

Yu, X., & Michael, H. A. (2019). Offshore pumping impacts onshore groundwater resources and land subsidence. Geophysical Research Letters, 46(5), 2553-2562.

Chen, H., Xue, Y., & Qiu, D. (2022). Numerical simulation of the land subsidence induced by groundwater mining. Cluster Computing, 1-10.

Wang, Y. Q., Wang, Z. F., & Cheng, W. C. (2019). A review on land subsidence caused by groundwater withdrawal in Xi’an, China. Bulletin of Engineering Geology and the Environment, 78, 2851-2863.

Doke, R., Kikugawa, G., & Itadera, K. (2020). Very local subsidence near the hot spring region in Hakone Volcano, Japan, inferred from InSAR time series analysis of ALOS/PALSAR data. Remote Sensing, 12(17), 2842.