frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 03 September 2025
DOI 10.3389/fpubh.2025.1672479

OPEN ACCESS

EDITED BY
Tong Wang,
Duke University, United States

REVIEWED BY
Ruiqi Zhang,
Massachusetts Institute of Technology, United States
Ran Tong,
The University of Texas at Dallas, United States
Hao Wu,
University of California San Francisco, United States
Ruobing Bai,
Alexion Pharmaceuticals, United States

*CORRESPONDENCE
Xiaoqing Peng
✉ xiaoqpeng5-c@my.cityu.edu.hk

RECEIVED 24 July 2025
ACCEPTED 21 August 2025
PUBLISHED 03 September 2025

CITATION
Zhu H and Peng X (2025) Decoding the association between health level and human settlements environment: a machine learning-driven provincial analysis in China. *Front. Public Health* 13:1672479. doi: 10.3389/fpubh.2025.1672479

Decoding the association between health level and human settlements environment: a machine learning-driven provincial analysis in China

Haidong Zhu and Xiaoqing Peng*

Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong, Hong Kong SAR, China

Background: Rapid urbanization in China has significantly reshaped the human settlement environment (HSE), bringing opportunities and challenges for public health. While existing studies have explored environmental-health relationships, most are confined to micro-level contexts, focus on single environmental dimensions, or assess specific diseases, thus lacking a comprehensive, macro-level understanding.

Objective: This study aims to assess the associations between population health level and multidimensional HSE features at the provincial level in China and uncover nonlinear relationships and interaction effects underlying the association between HSE and population health level.

Methods: Using panel data from 31 Chinese provinces spanning 2012 to 2022, a composite Health Level Index (HLI) was constructed based on four core health indicators using the Entropy-TOPSIS method. 19 HSE indicators covering five dimensions—ecological environment, living environment, infrastructure, public services, and sustainable environment—were selected as explanatory variables. The study employed the XGBoost machine learning algorithm to model the relationship between HSE and HLI. SHAP values and Partial Dependence Plots (PDPs) were used to interpret feature importance, nonlinear relationships, threshold values, and interaction effects.

Results: XGBoost outperformed all benchmark models, confirming its strong predictive capacity. SHAP analysis identified six key features—number of medical institution beds (NMIB), urbanization rate (UR), mobile phone penetration rate (MPPR), road area per capita (RAPC), population density (PD), and urban gas penetration rate (UGPR)—as the most influential factors. Nonlinear relationships and threshold effects were observed between key features and population health level. PDP plots further revealed that optimal health levels are typically associated with high UR, high MPPR, high RAPC, and moderate NMIB, underscoring the importance of structural synergy over isolated infrastructure expansion.

Conclusion: This study provides robust evidence that the relationship between HSE and health is nonlinear, multidimensional, and highly interactive. Effective urban health governance requires coordinated development of urbanization, digital infrastructure, and public services, along with rational healthcare resource allocation. The findings offer actionable insights for health-oriented urban planning and policy formulation in rapidly urbanizing regions.

KEYWORDS

health level, human settlement environment, machine learning, XGBoost, Shapley additive explanations

1 Introduction

Urbanization has reshaped the global HSE, especially in fast-growing economies such as China. As the world's largest developing country, China's urbanization level has risen rapidly over the decades, with a total of 694 cities at the end of 2023, compared to 129 cities at the start of the new China, according to the National Bureau of Statistics of China. By the end of 2023, the urbanization rate of China's resident population was 66.16%, compared to 10.64% at the end of 1949 (1).

The HSE refers to the living spaces inhabited by human populations. It is a geographical space closely associated with human survival and serves as a primary arena where humans utilize and transform nature (2). From a research perspective, the HSE is a multidimensional system, and the specific dimensions involved often vary depending on the research perspective. For example, some studies have evaluated the suitability of human settlements by integrating factors such as economic vitality, public services, infrastructure, topography, and climate conditions (3–5). Other scholars have focused on sustainability by incorporating elements like transportation, cultural resources, and living conditions into their assessments (6, 7). Additionally, certain studies have assessed the vulnerability of human settlements through dimensions such as the natural environment, social environment, and residential conditions (8, 9). At present, there is no universally agreed-upon set of evaluation indicators for human settlements; instead, indicator selection is typically determined by the specific research objectives, theoretical framework, and data availability.

At the same time of rapid urbanization, the contradiction between HSE and residents' health is becoming more and more prominent. It has been shown that large-scale urban expansion has led to the shrinkage of green space (10), the intensification of the heat island effect (11), and the spread of air and water pollution (12, 13). These changes directly or indirectly impact the physical or mental health of the population, for example, leading to an increase in the incidence of respiratory diseases, chronic diseases (14–16), and psychological problems such as anxiety and depression, among others (17). Assessing the extent to which the HSE is associated with population health level and exploring the predictive relationships between them is becoming a hotspot of interest in the fields of health, environment, and urban studies.

Existing researches on environment and health risk is abundant, but in terms of research scale, are mainly focus on community, urban environment (18, 19), lack of research in macro scale such as provincial, national, and so on, and in terms of selection of environmental factors, are mainly focus on single social environment, natural and built environments (20–22), lack of comprehensive exploration of multidimensional environmental factors, and in terms of health impacts, are limited to focusing on the environmental impacts on the risks of a specific disease (23), and lack of focus on the overall health of the population. To address these gaps, this study aims to examine the relationship between multidimensional HSE factors and the overall health level of the population at the provincial scale in China, using interpretable machine learning techniques to uncover nonlinear relationships and interaction effects.

2 Data and methods

2.1 Data

The dataset of this paper is divided into two parts. In the first part, referring to previous studies (24–26), four indicators were selected to represent the health level of the population: incidence of class A and B notifiable infectious diseases, mortality of class A and B notifiable infectious diseases, human mortality, and average life expectancy. In China, notifiable infectious diseases are categorized into three classes—A, B, and C—according to their potential threat to public health, with severity decreasing from A to C; classes A and B, which together encompass 29 diseases (2 in class A and 27 in class B), include the most serious infectious diseases and are therefore widely adopted as core measures of population disease burden (27). Together with human mortality, these disease-related indicators capture both the prevalence and fatality of major health threats, while average life expectancy provides a broader perspective on long-term population well-being, reflecting the cumulative effects of healthcare quality, living conditions, and social development. This combination of short-term disease burden and long-term health outcomes has been extensively used in health evaluation studies (28), and the data are consistently available from official statistical sources, ensuring reliability and comparability across regions. Based on these indicators, the Entropy–TOPSIS method was applied to construct the HLI, which integrates multiple dimensions of health into a single, comprehensive measure. Details of the four selected indicators are presented in Table 1. In the second part, according to the needs of the study, partly referring to the indicators used in previous studies (29–31), the HSE was divided into five dimensions: ecological environment, living environment, infrastructure conditions, public service, and sustainable environment, and 19 secondary indicators were selected. All HSE data can be directly obtained from the aforementioned public sources, except for four indicators—RAPC, number of sanitation vehicles per 10,000 population (NSV), PD,

TABLE 1 Indicators related to the level of health of the population.

Indicator	Unit	Indicator Attribute
Incidence of Class A and B notifiable infectious diseases	1/100,000	–
Mortality of Class A and B notifiable infectious diseases	1/100,000	–
Human mortality	%	–
Average life expectancy	Years	+

TABLE 2 HSE indicator system.

Primary indicator	Secondary indicator	Unit
Ecological environment	*Per capita* park green area (PCPGA)	m^2
	Chemical oxygen demand emissions (CODE)	10^4 tons
	Sulfur dioxide emissions (SO_2 Emissions)	10^4 tons
	Number of sanitation vehicles per 10,000 population (NSV)	Units
Living environment	Urban water penetration rate (UWPR)	%
	Urban gas penetration rate (UGPR)	%
	Population density (PD)	Persons/km^2
	Urbanization rate (UR)	%
Infrastructure conditions	Number of public toilets per 10,000 people (NPT)	Units
	Road area per capita (RAPC)	m^2
	Number of public transportation vehicles per 10,000 People(NPTV)	Standard units
	Mobile phone penetration rate (MPPR)	Units/100 persons
Public service	Number of higher education students per 100,000 people (NHES)	Persons
	Public library floor area per 10,000 population (PLFA)	m^2
	Number of medical institution beds per 10,000 People (NMIB)	Units
	Population served per postal service outlet (PSPSO)	10^4 persons
Sustainable environment	Daily urban sewage treatment capacity (DUSTC)	10^4 m^3
	Local financial environmental protection expenditure (LFEPE)	10^8 yuan
	Per capita daily domestic water consumption (PCDDWC)	Liters

and UR—which were derived through calculation. The calculation methods for these four indicators are presented in Equations 1–4. The details of HSE indicators are shown in Table 2.

The first part of the data comes from the China Health Statistics Yearbook 2013–2023, where the missing years of average life expectancy are filled in by linear interpolation, a method chosen based on relevant studies supporting the linear growth trend of life expectancy (32, 33), thus ensuring the scientific validity and completeness of the data. In total, 279 provincial-level data points on average life expectancy were supplemented. The second part of the data was obtained from the National Bureau of Statistics of China, China Statistical Yearbook 2013–2023, and Statistical Bulletins of Chinese provinces. The final dataset consists of these two parts of data, covering 31 provinces in China for the period 2012–2022. This yields a balanced panel of 341 province-year observations, with a total of 6,820 variable observations used in the analysis. The descriptive statistics of the variables are shown in Table 3.

$$\mathrm{RAPC} = \frac{\text{Total Road Area}}{\text{Resident Population at Year} - \text{End}} \quad (1)$$

$$\mathrm{NSV} = \frac{\text{Number of Sanitation Vehicles}}{\text{Resident Population at Year} - \text{End}} \times 10{,}000 \quad (2)$$

$$\mathrm{PD} = \frac{\text{Resident Population at Year} - \text{End}}{\text{Provincial Administrative Area}} \quad (3)$$

$$\mathrm{UR} = \frac{\text{Urban Population}}{\text{Resident Population at Year} - \text{End}} \quad (4)$$

2.2 Method

2.2.1 HLI construction based on entropy-TOPSIS method

Various evaluation models, such as Fuzzy Comprehensive Evaluation (34), the Analytical Hierarchy Process (AHP) (35), and the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) (36), have been utilized in recent studies. To scientifically quantify the population health level in each province in China, this study introduces the entropy weight method combined with the TOPSIS evaluation model to construct the HLI. While standard statistical techniques like Principal Component Analysis (PCA) or Factor Analysis could create weighted composites based on covariance structures (37), our study employs the Entropy-TOPSIS approach for several key advantages: 1. The entropy weight method objectively determines indicator weights based on information content rather than subjective expert judgment (38), avoiding potential bias inherent in AHP or equal weighting schemes; 2. Unlike PCA, which may lose interpretability through linear combinations of variables, entropy weighting preserves the original meaning of each health indicator; 3. TOPSIS provides intuitive relative rankings by measuring proximity to ideal solutions (39), making results more accessible for policy interpretation compared to factor scores; 4. The entropy weight method is used to avoid subjective bias, and the TOPSIS method is used to measure the relative closeness of the samples to the ideal solution, resulting in the formation of the HLI, which is both scientific and comparable. Prior to constructing the composite index, a correlation analysis was conducted on the four variables to ensure their appropriateness for inclusion, with the results presented in Figure 1. The results showed that the absolute values of the correlation coefficients between the variables were all below 0.6, indicating no

TABLE 3 Descriptive statistics of variables.

Variable	Sample size	Mean	Std. dev.	Minimum	Maximum
PCPGA	341	13.47	2.84	5.85	22.84
CODE	341	56.85	49.59	1.76	192.12
SO_2 emissions	341	33.05	35.35	0.11	174.88
NSV	341	1.95	1.64	0.12	15.02
UWPR	341	97.87	3.03	67.57	100.00
UGPR	341	93.78	9.22	29.79	100.00
PD	341	460.36	701.53	2.57	3925.87
UR	341	59.80	12.68	22.86	89.58
NPT	341	3.18	1.26	0.77	9.35
RAPC	341	5.97	2.29	1.14	13.71
NPTV	341	12.73	3.01	5.63	26.55
MPPR	341	104.63	24.07	57.30	189.46
NHES	341	2774.84	855.97	1133.00	5534.00
PLFA	341	118.51	48.23	45.66	335.80
NMIB	341	56.51	11.35	27.15	84.31
PSPSO	341	0.72	0.44	0.15	2.77
DUSTC	341	573.09	497.78	5.00	2971.30
LFEPE	341	155.81	100.91	17.21	747.44
PCDDWC	341	174.24	49.19	91.12	403.62
HLI	341	0.60	0.11	0.36	0.89

severe multicollinearity and supporting their suitability for constructing the composite index. The specific steps of the Entropy-TOPSIS method are as follows:

Step 1: Matrix construction of raw data.

Assuming the research object contains m samples ($i = 1, 2, \cdots, m$), 4 core health indicators (Table 1) are selected to form the evaluation system, and the raw data matrix X is defined as:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & x_{m4} \end{bmatrix} \tag{5}$$

Where: x_{ij} denotes the raw value of the j-th indicator for the i-th sample.

Step 2: Data standardization.

To ensure comparability and preserve the directional meaning of each indicator, positive and negative indicators are normalized separately.

(1) For positive indicators:

$$x_{ij}^{'} = \frac{x_{ij} - \min\left(x_j\right)}{\max\left(x_j\right) - \min\left(x_j\right)} \tag{6}$$

(2) For negative indicators:

$$x_{ij}^{'} = \frac{\max\left(x_j\right) - x_{ij}}{\max\left(x_j\right) - \min\left(x_j\right)} \tag{7}$$

The processing yields a normalized matrix X'

$$X' = \begin{bmatrix} x_{11}^{'} & x_{12}^{'} & x_{13}^{'} & x_{14}^{'} \\ x_{21}^{'} & x_{22}^{'} & x_{23}^{'} & x_{24}^{'} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1}^{'} & x_{m2}^{'} & x_{m3}^{'} & x_{m4}^{'} \end{bmatrix} \tag{8}$$

Step 3: Calculation of indicator weights based on the entropy weight method.

(1) Calculate the proportion of j-th indicator of the i-th sample:

$$r_{ij} = \frac{x_{ij}^{'}}{\sum_{i=1}^{m} x_{ij}^{'}} \tag{9}$$

(2) Calculate the information entropy e_j for the jth indicator:

$$e_j = -\frac{1}{\ln m} \sum_{i=1}^{m} r_{ij} \ln r_{ij} \tag{10}$$

(3) Calculation of the indicator coefficient of variation d_j:

$$d_j = 1 - e_j \tag{11}$$

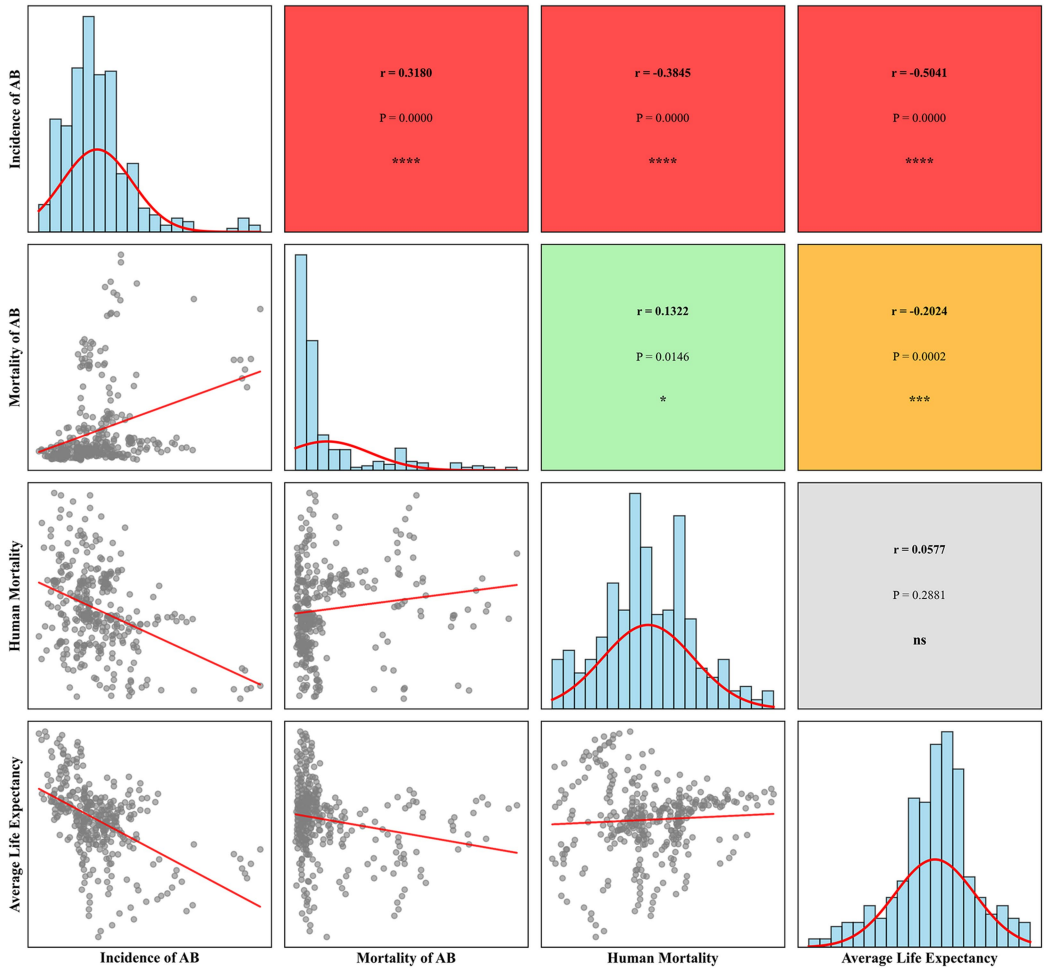

FIGURE 1
Correlation coefficient matrix of health indicators.

(4) Determine the entropy weight w_j:

$$w_j = \frac{d_j}{\sum_{k=1}^{4} d_k} \tag{12}$$

The weight results are shown in Figure 2.

Step 4: TOPSIS method.

(1) Construct the weighted normalization matrix V:

$$V = \begin{bmatrix} w_1 x_{11}' & w_2 x_{12}' & w_3 x_{13}' & w_4 x_{14}' \\ w_1 x_{21}' & w_2 x_{22}' & w_3 x_{23}' & w_4 x_{24}' \\ \vdots & \vdots & \vdots & \vdots \\ w_1 x_{m1}' & w_2 x_{m2}' & w_3 x_{m3}' & w_4 x_{m4}' \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & v_{13} & v_{14} \\ v_{21} & v_{22} & v_{23} & v_{24} \\ \vdots & \vdots & \vdots & \vdots \\ v_{m1} & v_{m2} & v_{m3} & v_{m4} \end{bmatrix} \tag{13}$$

(2) Determine the ideal solution:

$$\begin{aligned} S_j^+ &= \max\left(v_{1j}, v_{2j}, \cdots, v_{mj}\right) \\ S_j^- &= \min\left(v_{1j}, v_{2j}, \cdots, v_{mj}\right) \end{aligned} \tag{14}$$

(3) Calculate the distance from the sample to the ideal solution:

$$\begin{aligned} D_i^+ &= \sqrt{\sum_{j=1}^{4}\left(v_{ij} - S_j^+\right)^2} \\ D_i^- &= \sqrt{\sum_{j=1}^{4}\left(v_{ij} - S_j^-\right)^2} \end{aligned} \tag{15}$$

(4) Calculate the closeness HLI_i

$$HLI_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{16}$$

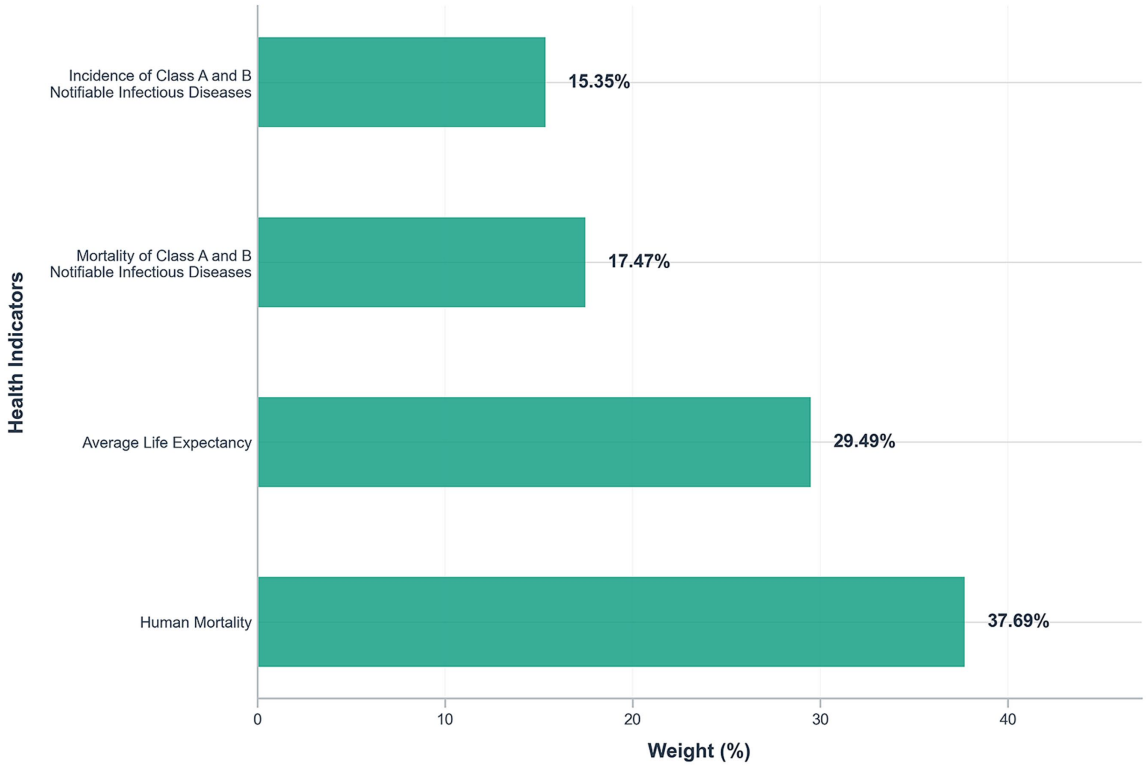

FIGURE 2
Weigh of health indicators.

$HLI_i \in [0,1]$ with larger values indicating a higher health level of the province's residents, and smaller values indicating a lower health level.

2.2.2 XGBoost model

XGBoost (eXtreme Gradient Boosting) is an optimized implementation of the Gradient Boosting Decision Tree (GBDT) algorithm, which builds a strong learner by integrating multiple weak learners to achieve highly accurate predictions. The algorithm performs well in classification and regression tasks dealing with structured data, and is particularly suitable for machine learning scenarios with high feature dimensionality and large sample sizes (40)

XGBoost builds an additive model composed of K regression trees to predict an output $\hat{y}_i$ for each sample i. The model prediction is defined as

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathrm{x}_i), \quad f_k \in \mathcal{F} \tag{17}$$

where $\mathrm{x}_i \in \mathbb{R}^d$ is the input feature vector and $\mathcal{F}$ is the space of regression trees. Each tree f_k maps x_i to a leaf score.

The objective function, minimized at the t-th boosting iteration, combines a convex loss function l and a regularization term Ω that controls tree complexity:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathrm{x}_i)\right) + \Omega(f_t) \tag{18}$$

Using a second-order Taylor expansion around $\hat{y}_i^{(t-1)}$, the objective approximates to

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\mathrm{x}_i) + \frac{1}{2} h_i f_t^2(\mathrm{x}_i) \right] + \Omega(f_t) \tag{19}$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ are the first and second derivatives of the loss. The regularization term typically includes the number of leaves T and leaf weights w_j:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{20}$$

This allows efficient tree structure optimization via greedy split finding and leaf weight calculation.

In this study, the XGBoost model was implemented in Python. The optimal parameters were obtained using a random search method, taking into account the computational efficiency (41). The dataset was divided into training and test data at a ratio of 8:2 using a fixed random seed (seed = 2025) to ensure reproducibility, with resampling conducted at the province-year level to maintain temporal and spatial independence and avoid data leakage across different provinces and years. The K-fold cross-validation method was used (the value of K was 5 in this study). This method involves

partitioning the training data into five equal subsets and iteratively using one subset for validation while training on the remaining four, a process repeated five times to generate a comprehensive assessment of the model's generalization performance and effectively avoid the overfitting problem of the model presented during the training process. The finalized key parameters for the XGBoost model were set as follows: learning rate 0.05; n_estimators 200; max_depth 3.

2.2.3 SHAP method

SHAP (SHapley Additive exPlanations) is based on the Shapley value theory from game theory, providing a unified interpretability framework for machine learning models (42). SHAP values provide a theoretically grounded measure of feature attribution by interpreting model output as an additive feature attribution:

$$\hat{y}_i = \phi_0 + \sum_{j=1}^{d} \phi_j x_{ij} \tag{21}$$

Where $\phi_0 = \mathbb{E}[\hat{y}]$ is the expected model output, and ϕ_j quantifies the contribution of feature j to the prediction for instance i.

Formally, the SHAP value ϕ_j is calculated as follows:

$$\phi_j = \sum_{S \subseteq \mathcal{D}\{j\}} \frac{|S|!(d-|S|-1)!}{d!} \left[f_{S\cup j}(x_{S\cup j}) - f_S(x_S) \right] \tag{22}$$

where $\mathcal{D} = \{1,\ldots,d\}$ is the full feature set, S is a subset of features excluding j, and $f_S(x_S)$ denotes the model output when only features in S are present. This weighted average of marginal contributions satisfies properties of local accuracy, consistency, and missingness, ensuring fair and interpretable explanations.

2.2.4 Model evaluation method

This study uses three statistical metrics to evaluate model prediction performance: coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The specific formulas (Equations 23–25) are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{23}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{24}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{25}$$

Where:

y_i represents the actual value for the i-th sample.

$\hat{y}_i$ represents the predicted value for the i-th sample.

$\bar{y}$ represents the mean of all actual sample values.

n represents the total number of samples.

3 Results and analysis

3.1 Model performance comparison

In this study, six machine learning algorithms—XGBoost, AdaBoost, Gradient Boosting Decision Trees (GBDT), LightGBM, Random Forest, and Lasso regression—were systematically evaluated and compared in terms of their predictive performance. The detailed numerical results are presented in Table 4, while the corresponding residual plots and fitting performance plots are illustrated in Figure 3, providing a visual representation of the models' predictive accuracy and residual distributions.

Overall, XGBoost demonstrated the most robust performance among all models across multiple evaluation metrics, highlighting its superior ability to capture complex, nonlinear relationships within the data. In particular, on the test set, XGBoost achieved a coefficient of determination (R^2) of 0.929, which was not only substantially higher than that of the conventional linear regression approach (Lasso regression, $R^2 = 0.824$) but also exceeded the performance of other ensemble-based methods such as AdaBoost ($R^2 = 0.910$) and GBDT ($R^2 = 0.917$). Furthermore, XGBoost attained the lowest root mean square error (RMSE = 0.033) and mean absolute error (MAE = 0.026) on the test dataset, underscoring its high predictive accuracy, minimal bias, and strong generalization capability.

The residual plots in Figure 3 further reinforce these findings, showing that the residuals of the XGBoost model are symmetrically distributed around zero, with no apparent heteroscedasticity or systematic patterns, indicating an adequate model fit and effective mitigation of overfitting. Similarly, the fitting performance plots depict strong linear correlations between the predicted and actual values for both the training and test sets, with the majority of points closely

TABLE 4 Model performance comparison.

Model	Training set R^2	Test set R^2	Training set RMSE	Test set RMSE	Training set MAE	Test set MAE
XGBoost	0.988	0.929	0.012	0.033	0.009	0.026
GBDT	0.983	0.917	0.014	0.035	0.011	0.029
Adaboost	0.974	0.910	0.018	0.037	0.015	0.031
LightGBM	0.941	0.881	0.026	0.042	0.021	0.036
Random forest	0.932	0.871	0.029	0.044	0.022	0.038
Lasso regression	0.792	0.824	0.050	0.052	0.041	0.043

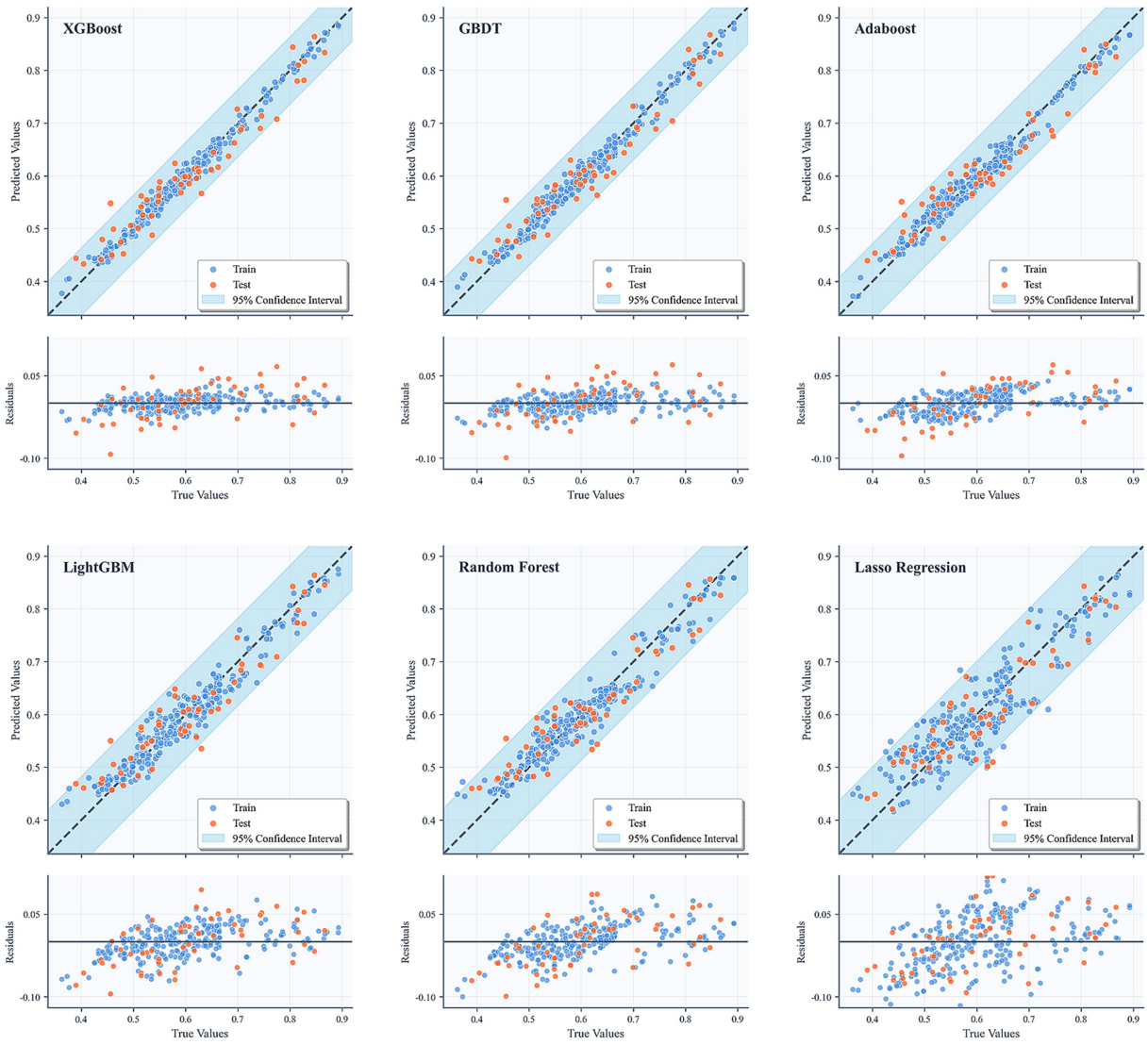

FIGURE 3
Performance of six models on the test and training sets.

aligning with the 1:1 diagonal line, which represents perfect prediction. The inclusion of 95% confidence intervals provides an additional layer of interpretability, offering reasonable uncertainty bounds and further validating the reliability and robustness of the model's predictions. Collectively, these results confirm that XGBoost not only delivers superior performance compared to traditional regression models and other ensemble methods but also maintains consistent accuracy and stability across different evaluation criteria, making it a compelling choice for predictive modeling in this context.

3.2 Feature importance analysis

SHAP values provide insights into both the direction (positive or negative) and the magnitude of each feature's contribution to model predictions, offering a quantitative basis for identifying key influencing factors. Based on the SHAP mean absolute values, the importance ranking and distribution of 19 features in predicting the health level of the population are illustrated in Figure 4, while the percentage contribution of each feature is shown in Figure 5.

The SHAP summary plot and corresponding feature importance percentages reveal significant heterogeneity in the associations between various HSE features and population health. Among these, the top six most influential features— NMIB:23.52%, UR:12.18%, MPPR:11.71, RAPC:7.93%, PD:6.93%, and UGPR:6.92%—contribute substantially more than others, indicating their central role in reflecting health level.

In the SHAP summary plot, NMIB displays a counterintuitive distribution: higher NMIB is predominantly associated with negative SHAP values, whereas lower NMIB corresponds more frequently with positive SHAP values. This pattern is consistently observed across the other five models, as shown in Figure 6, providing further support for our finding. From the perspective of health demand, the number of hospital beds reflects the health status and healthcare needs of local populations. Regions with a high NMIB often experience greater disease burdens, higher prevalence of chronic illnesses, or more

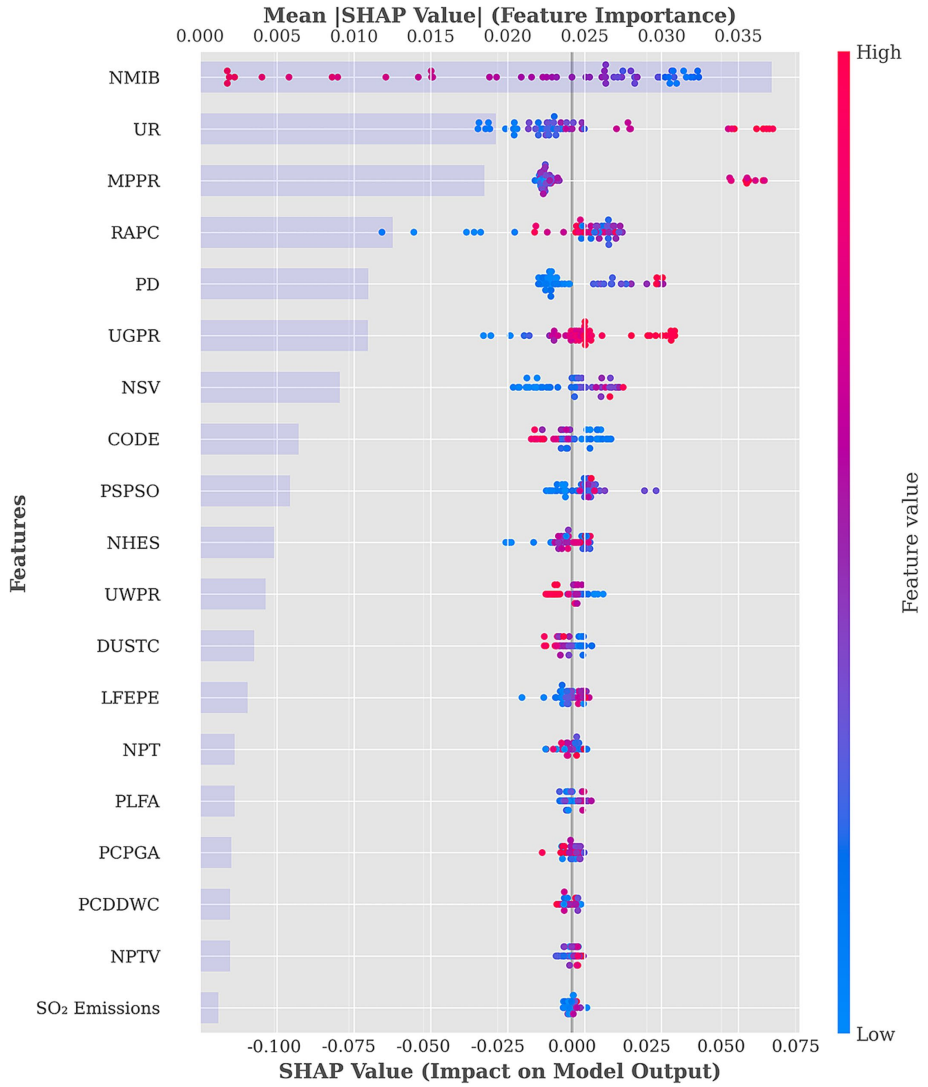

FIGURE 4
Feature summary and importance plot.

advanced population aging. These areas require more hospital beds to meet elevated inpatient service demands. Consequently, a high NMIB may signal relatively poor population health level—an observation similar to prior findings (43, 44). From a healthcare service model perspective, modern systems increasingly prioritize disease prevention and outpatient care, aiming to reduce unnecessary hospitalizations. Healthier regions tend to have more robust public health infrastructures and advanced medical technologies, enabling effective prevention and outpatient management, which in turn reduces inpatient demand. Thus, such regions may exhibit lower NMIB while maintaining a higher overall health level. Population structure is another relevant factor. Areas with older populations—who typically require more frequent inpatient care—demand greater hospital bed capacity. Conversely, younger populations have lower hospitalization needs, reflected in lower NMIB. This 'demand response effect' suggests that the evaluation of healthcare resource allocation should consider health needs and demographic structures rather than assuming 'more is better'. Optimal resource distribution should ensure basic needs are met while improving efficiency and outcomes through smarter planning and service innovation (45).

UR demonstrates a clear gradient in the SHAP value distribution, transitioning from negative SHAP values at low urbanization levels to positive values at higher levels. This pattern reflects the progressive health-promoting association of urbanization. Urbanized areas typically benefit from more comprehensive healthcare facilities and public service systems. Urbanization also brings improved sanitation infrastructure—such as water supply, sewage, and waste management—significantly enhancing living conditions and health security. As noted by Ngounou, Oumbe (46), urbanization also positively affects education. Residents in highly urbanized areas are more likely to access health education, disease prevention information, and modern health concepts. Enhanced health literacy fosters healthier lifestyles, greater self-care awareness, and better disease prevention, contributing to long-term health improvements. Urban environments also offer cultural, recreational, and fitness facilities that promote physical and mental well-being (47).

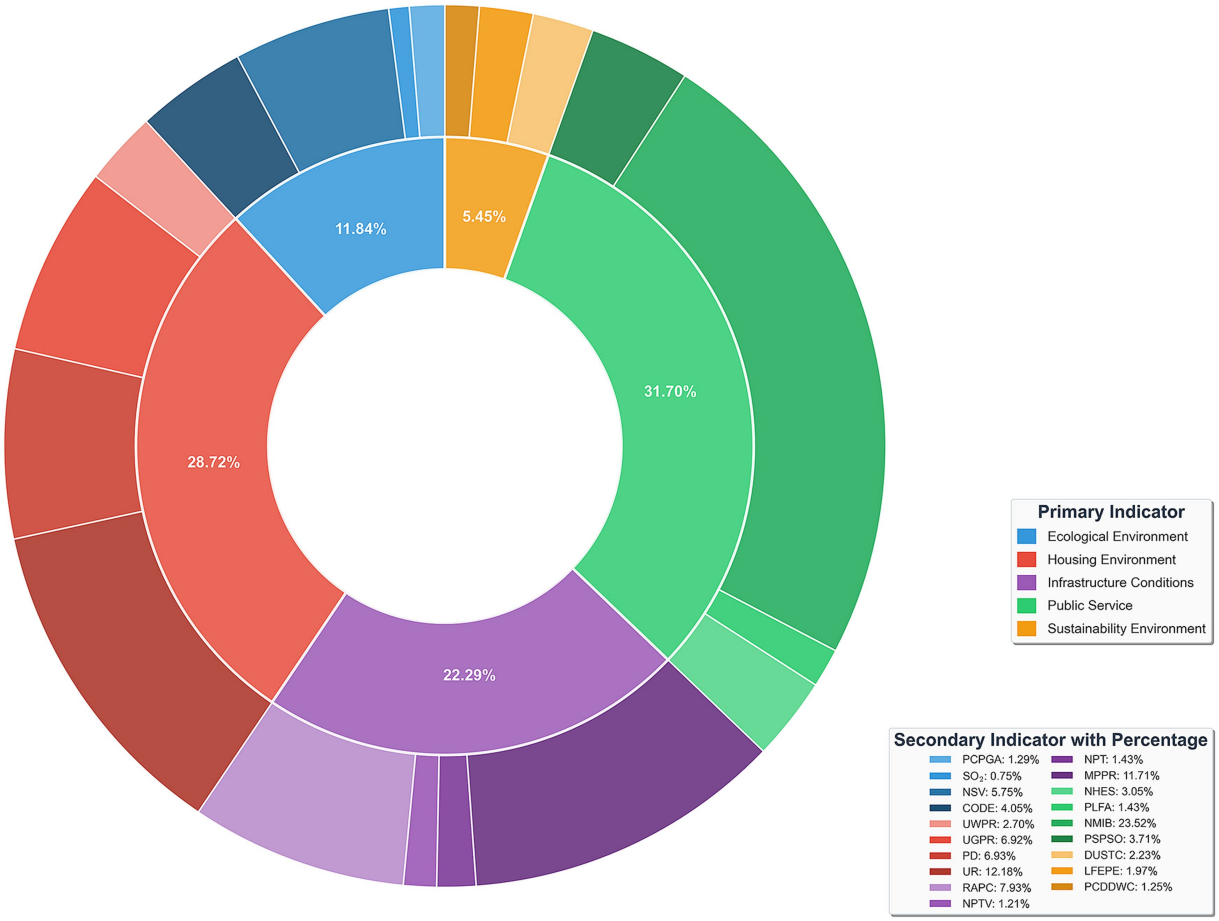

FIGURE 5
Percentage of feature importance.

MPPR exhibits a clear polarization in its SHAP value distribution: low MPPR clusters around negative SHAP values, while high MPPR align with positive values. This underscores the role of the digital divide in health. Improved mobile phone penetration significantly enhances residents' ability to access health-related information (48). In areas with high MPPR, residents can easily obtain disease prevention guidance, health behavior tips, and medical service information, contributing directly to better health literacy and behavioral improvements. In underserved regions, mobile internet acts as a vital supplement to traditional health education. From a healthcare access perspective, MPPR supports the development of digital healthcare services (49). The proliferation of telemedicine, online consultations, e-prescriptions, and mobile health apps helps mitigate healthcare resource imbalances, particularly enhancing service accessibility in under-resourced areas. Socially, mobile phones facilitate community engagement and social integration. Digital platforms enable participation in local activities, access to social support, and maintenance of interpersonal relationships—all of which are beneficial to mental health and overall well-being. However, it is important to acknowledge the potential downsides of mobile technology (50). In areas with low MPPR, limited information access and service availability may hinder health improvement, reflecting the adverse effects of the digital divide.

RAPC demonstrates a complex, nonlinear pattern in the SHAP value distribution. High RAPC values appear at both positive and negative values of the SHAP value spectrum, while moderate values tend to cluster around positive SHAP values. This indicates the multifaceted associations between transport infrastructure and health level. Access to transportation is closely linked to healthcare accessibility (51). Efficient transportation networks reduce commuting burden and emergency response times. In medical emergencies, accessible roads improve ambulance response and increase survival rates. However, excessive road infrastructure may also lead to adverse health outcomes. Densely developed road networks can increase traffic volume, air pollution, noise, and accident risks (52–54). In urban cores, expanded road areas may reduce green space and public recreational areas, diminishing environmental livability. In addition, overreliance on motorized transport may discourage physical activity such as walking or cycling, negatively impacting physical and mental health (55). The nonlinear effects of RAPC emphasize the importance of urban planning. Health-optimized transport infrastructure should balance mobility, environmental quality, and quality of life through better network design, investment in public transport, and the promotion of green mobility.

PD exhibits a clear gradient in SHAP value distribution: low PD is associated with negative SHAP values, while high PD is associated with positive values. This pattern suggests that moderate population agglomeration contributes positively to health levels. From a public service economy of scale perspective, higher PD facilitates more efficient allocation of healthcare, education, and cultural resources.

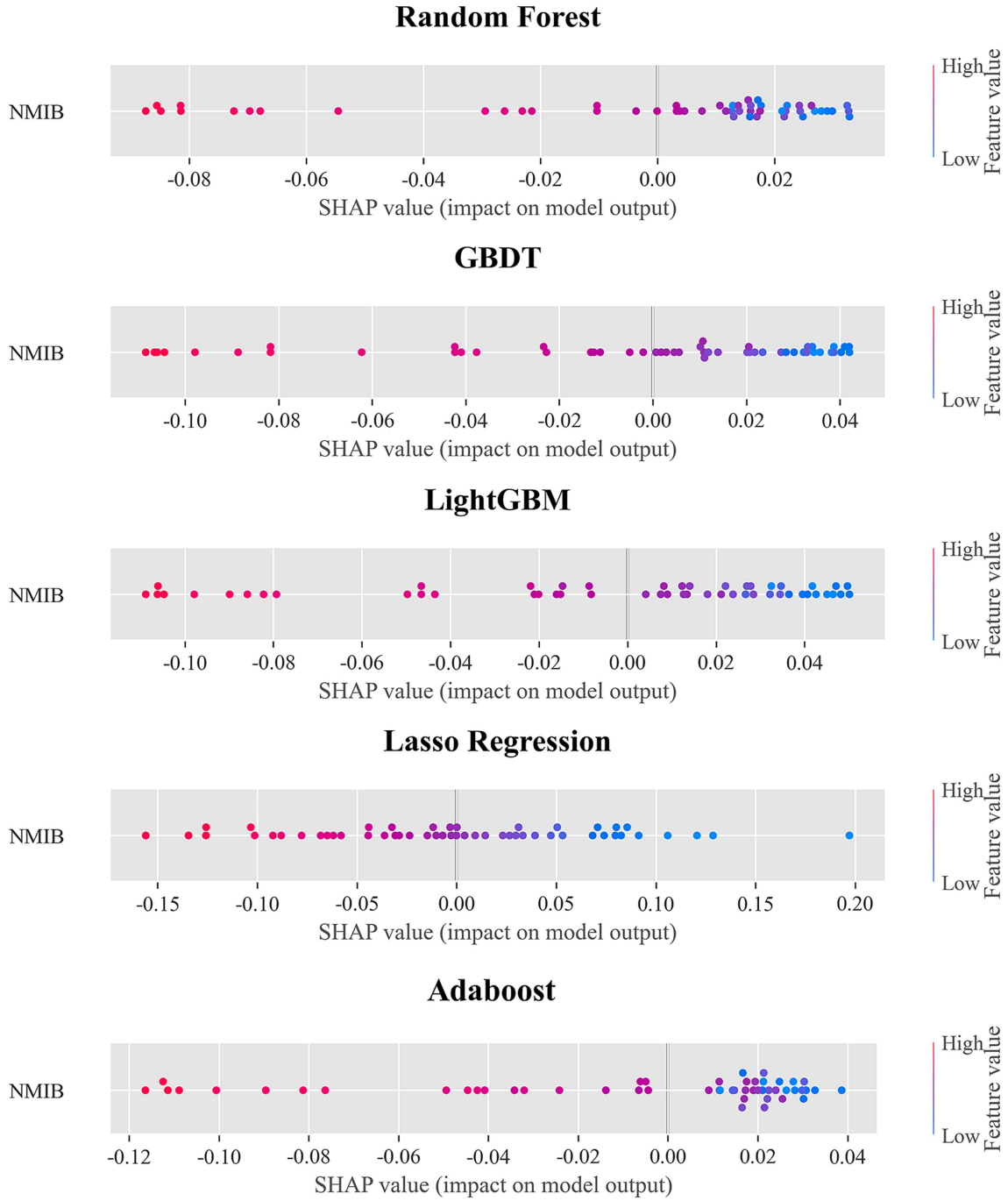

FIGURE 6
SHAP summary plot for NMIB in other models.

Densely populated areas often offer more specialized and diverse services, including segmented healthcare services, better educational access, and greater availability of sports and cultural amenities (56). From the perspective of social support networks, higher population density is significantly associated with an increase in social support (57). In communities with relatively concentrated populations, social interactions among residents are more frequent, community cohesion is stronger, and an effective social support system can be formed. This kind of social support network plays a significant role in maintaining mental health, disease prevention, and promoting healthy behaviors. However, excessive population density may result in environmental degradation (58). Thus, the analysis highlights the positive associations between moderate population agglomeration and health level rather than uniformly endorsing high-density development.

UGPR, an important indicator of clean energy adoption, shows a distinctly positive association with health level in the SHAP analysis. High UGPR values correspond to positive SHAP values, whereas low UGPR is associated with negative SHAP values, underscoring the significance of energy structure optimization for public health. Gas coverage's contribution to health predictions is primarily associated

with improvements in air quality. Compared to traditional coal combustion, natural gas—being a cleaner fossil fuel—generates significantly fewer pollutants such as particulate matter, sulfur dioxide, and nitrogen oxides (59). Regions with high gas penetration have lower air pollution emissions and better air quality, thereby reducing the incidence of respiratory diseases. This is particularly important in northern regions where gas has replaced coal for winter heating. From an indoor air quality perspective, widespread gas use significantly improves the domestic environment. Traditional coal-based heating and cooking generate substantial indoor pollutants such as carbon monoxide and sulfur dioxide. Clean gas combustion reduces these risks dramatically (60), lowering residents' exposure to hazardous indoor air. UGPR also reflects the modernization of urban infrastructure. High gas coverage requires comprehensive pipeline systems, safety protocols, and user-friendly service mechanisms, indicating advanced urban governance and public service quality.

3.3 Non-linear relationship and threshold effects analysis

To elucidate the complex nonlinear associations and potential threshold effects between the top six important features and HLI, this study employed locally weighted scatterplot smoothing (LOWESS) on SHAP value scatter plots to derive smoothed fitting curves. A smoothing span of 0.3 was determined through iterative experimentation and visual evaluation to balance the risks of overfitting and underfitting. A locally linear fitting method was adopted to enhance the robustness of the estimation. The threshold for each feature was defined as the point on the horizontal axis at which the fitted curve intersected the reference line corresponding to a SHAP value of zero. The uncertainty associated with these thresholds was quantified by constructing 95% confidence intervals using a bootstrap resampling procedure. The plots are shown in Figure 7.

Figure 7A reveals a distinct nonlinear relationship between NMIB and population health level, with a critical turning point at 60.21 (95% CI: 59.60–60.73) beds per 10,000 people. Below this threshold, increases in NMIB are associated with positive SHAP values, indicating that a higher availability of hospital beds contributes significantly to better health levels. This aligns with expectations, as adequate inpatient capacity ensures timely and effective treatment, thereby reducing morbidity and mortality rates. However, once NMIB exceeds approximately 60 beds per 10,000, SHAP values become negative. This phenomenon may reflect inefficient utilization of medical resources in certain provinces (61), or it may signal more severe public health challenges that require disproportionately greater healthcare infrastructure to manage.

Figure 7B illustrates the complex associations between urbanization and population health level, with a threshold at 65.73% (95% CI: 61.99–69.00). At a low urbanization level, SHAP values are negative, suggesting that early stages of urbanization are associated with a poorer health level. This may be correlated with increased environmental pollution, lifestyle shifts, and heightened social stress associated with rapid urban transition. Importantly, this negative association diminishes as urbanization progresses. After surpassing the threshold range, the SHAP values turn positive, indicating a reversal in the association. At this stage, the benefits of urban development—such as centralized medical resources, improved infrastructure, and higher educational attainment—begin to outweigh earlier disadvantages. Highly urbanized areas typically offer superior health systems, robust public health infrastructure, and greater health awareness, collectively contributing to improved population health levels. Overall, urbanization has played a positive role in the health level of the population, which is consistent with previous studies (62, 63).

Figure 7C identifies a distinct threshold effect at 123.21 (95% CI: 116.24–145.08) units/100 persons. Below this level, the association between increasing MPPR and health level is limited, with SHAP values remaining relatively low. However, once MPPR exceeds this threshold, its positive contribution to health predictions becomes pronounced and stabilizes. This suggests a digital threshold pattern: when mobile connectivity reaches a certain saturation point, digital health services—such as telemedicine, health monitoring, and access to medical information—become widespread. The availability of these services is associated with improved healthcare accessibility and efficiency.

Figure 7D highlights an early threshold effect for RAPC at 3.99 (95% CI: 3.75–4.25) m^2 per person. Below this level, increases in RAPC are associated with negative SHAP values, possibly due to negative externalities such as pollution, noise, and disruptions linked to early-stage road construction. Once RAPC surpasses the threshold range, the relationship turns positive and remains relatively stable. This indicates that a certain level of transport infrastructure improves access to healthcare services and supports health-related mobility. However, it also underscores the need to balance improved accessibility with potential environmental and social costs of overdevelopment.

Figure 7E reveals a threshold effect of PD on health, with a key turning point at 517.38 (95% CI: 472.23–609.68) persons/km^2, after which the health level increases until reaching a plateau at approximately 1,250 persons/km^2. In the low-density phase, increases in PD slightly reduce health level, as reflected in negative SHAP values. Beyond the threshold range, the association turns significantly positive, indicating that economies of scale associated with population concentration begin to show positive correlations with health level. Moderate population density facilitates efficient distribution of healthcare resources, scaled public health services, and stronger social networks. At around 1,250 persons/km^2, the fitted curve plateaus, and SHAP values stabilize at a high level, indicating a diminishing marginal effect of PD on health. This suggests an optimal density range where health benefits from population agglomeration. Several mechanisms may explain this plateau effect. Firstly, the allocation of medical resources and the public health service system in high PD areas are already relatively well-established. Further increases in population density are unlikely to yield significant marginal improvements in public health levels. Secondly, excessive PD may give rise to a range of negative factors, including intensified environmental pollution (64), heightened perception of stress (65), and an increased risk of infectious disease transmission (66). These adverse effects may offset some of the health benefits associated with population agglomeration.

Figure 7F reveals a high threshold of UGPR at 95.18 (95% CI: 94.18–96.50) %. Before reaching this threshold, UGPR has a slight negative trend in SHAP values. However, once gas penetration exceeds the threshold range, its contribution to health predictions sharply turns positive. This inflection likely reflects network and quality effects of gas infrastructure.

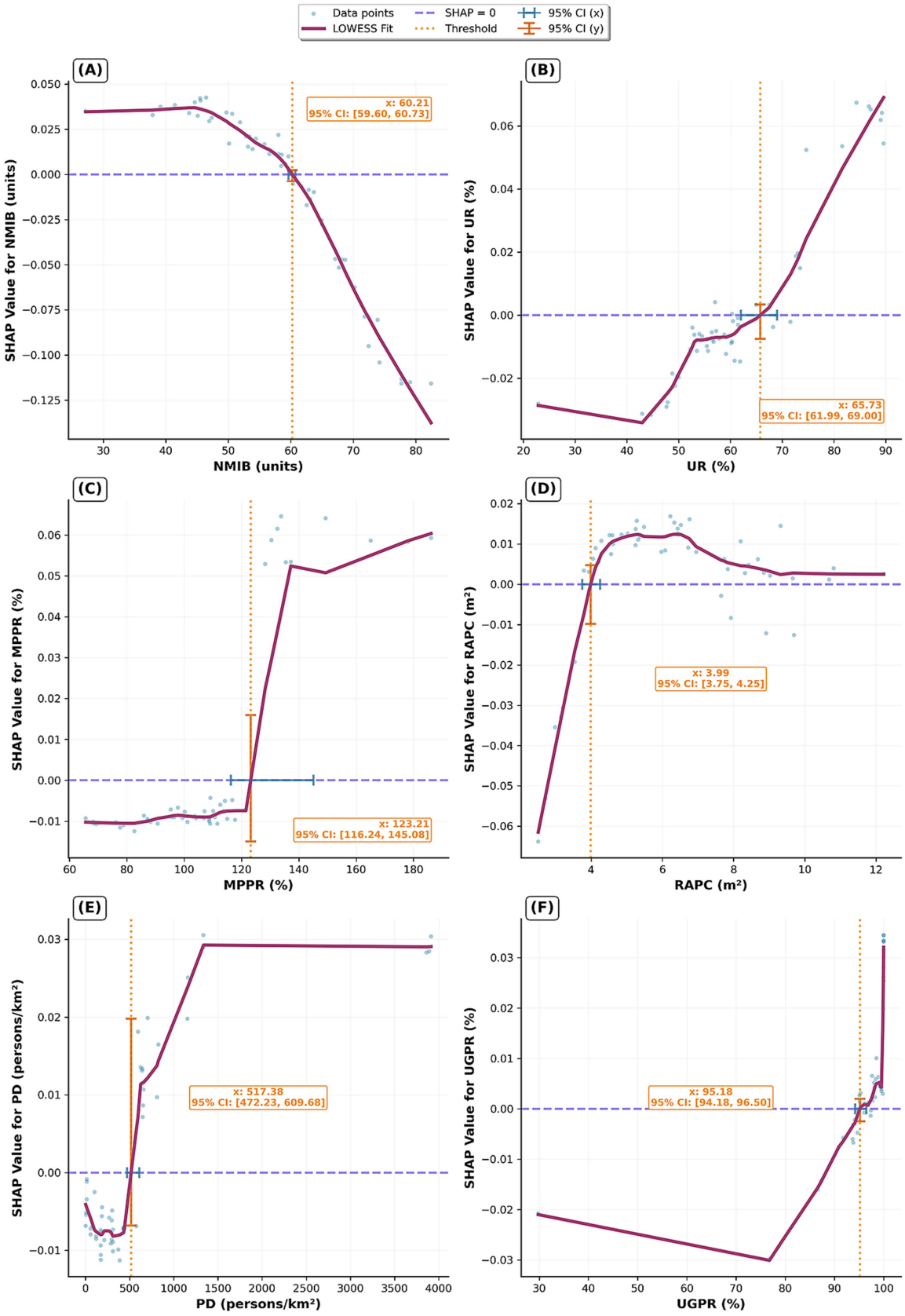

FIGURE 7
Nonlinear impacts and threshold effects of the HSE on HLI. **(A)** NMIB; **(B)** UR; **(C)** MPPR; **(D)** RAPC; **(E)** PD; **(F)** UGPR.

3.4 Analysis of interaction effects among HSE features

The PDP plots (Figure 8) reveal complex interactive effects of NMIB, UR, MPPR, and RAPC on HLI. High HLI values are typically observed under conditions of high UR, high MPPR, and high RAPC. However, NMIB does not exhibit a simple positive relationship with HLI; instead, a reverse association is identified.

Specifically, in regions with relatively low NMIB but high UR (Figure 8A), HLI tends to be higher. This suggests that urbanization is associated with higher population health levels, potentially linked to improved living conditions, healthcare accessibility, public health infrastructure, and healthier lifestyles. A lower NMIB may indicate reduced health pressure or more efficient medical resource utilization in these areas.

In Figure 8B, higher HLI values are concentrated in regions with high MPPR and moderate-to-low NMIB, implying that robust digital infrastructure is associated with more efficient health management and service delivery, which may reduce dependency on large numbers of hospital beds. Conversely, in areas with low MPPR, even with a high NMIB, HLI remains suboptimal. This indicates that merely expanding healthcare infrastructure is not necessarily associated with higher health levels without adequate digital support.

Figure 8C shows that high HLI values primarily occur in regions with high RAPC and moderate-to-low NMIB. This demonstrates that sufficient transportation infrastructure is associated with improved medical accessibility and emergency responsiveness, which may be linked to lower health risks and reduced reliance on hospital beds. In contrast, limited road infrastructure is associated with traffic congestion and delayed emergency response, potentially constraining the effectiveness of medical resources, regardless of bed availability.

Moreover, UR demonstrates significant synergistic interactions with MPPR and RAPC (Figures 8D–F). The combination of high UR and high MPPR is associated with higher HLI, potentially reflecting complementary effects of health information access and dissemination. Similarly, high UR paired with high RAPC is associated with higher health levels, potentially reflecting reduced urban congestion and medical service delays. The synergy between digitalization (MPPR) and physical infrastructure (RAPC) is also evident, where improvements in both simultaneously elevate the health level. High MPPR enhances the responsiveness and reach of health services, while high RAPC ensures spatial accessibility for medical resource allocation.

In summary, the PDP plots illustrate that NMIB does not simply reflect the adequacy of healthcare resources but also serves as a composite indicator of urban health risks, disease burden, and healthcare system efficiency. Improvements in HLI correlate with the coordinated contributions of multiple environmental factors. Solely expanding hospital capacity is not necessarily associated with higher health levels; it may reflect inefficiencies or reactive health governance. Optimal HLI levels are generally found in conditions characterized by high UR, strong digital infrastructure (MPPR), Well-developed infrastructure (RAPC), and moderate hospital bed supply. This highlights the importance of structural optimization and systemic synergy in urban health governance. Future health planning should

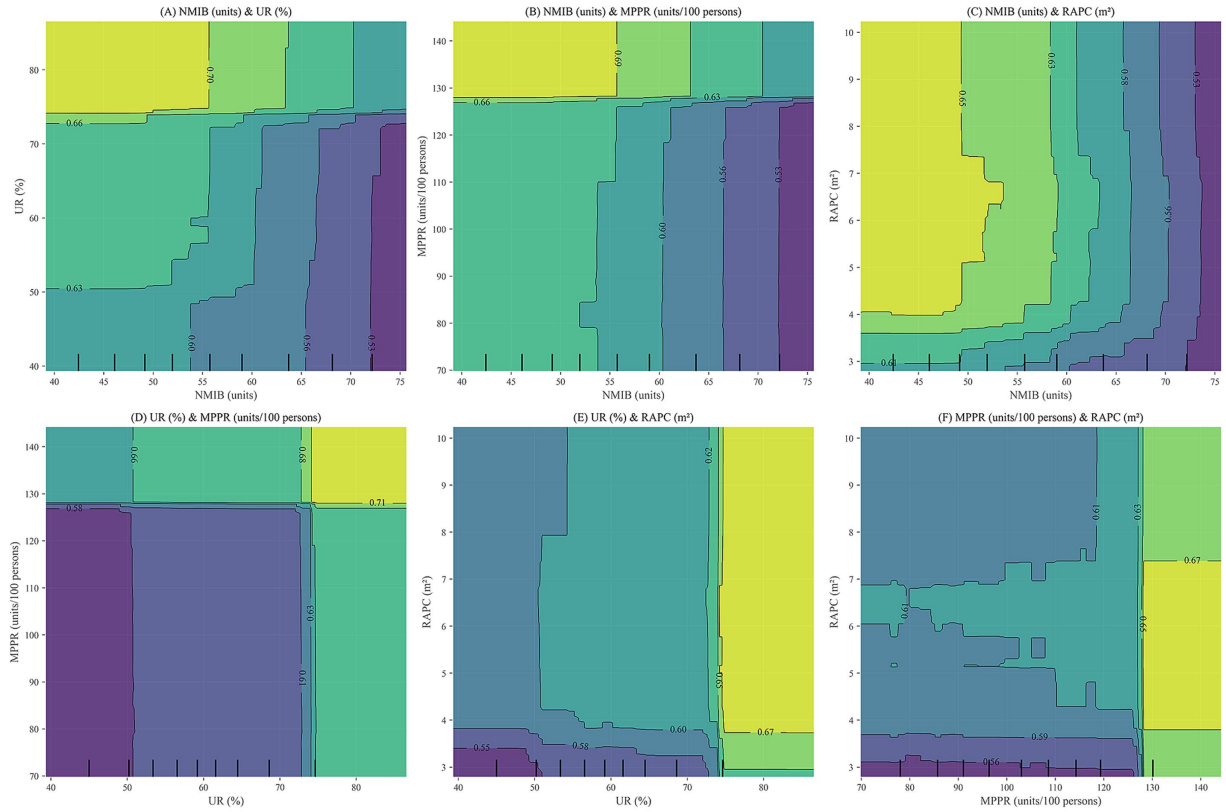

FIGURE 8
PDP plots for NMIB, UR, MPPR, and RAPC. **(A)** NMIB & UR; **(B)** NMIB & MPPR; **(C)** NMIB & RAPC; **(D)** UR & MPPR; **(E)** UR & RAPC; **(F)** MPPR & RAPC.

focus on the integrated development of urbanization, informatization, and infrastructure while improving the operational efficiency of healthcare systems and rational allocation of hospital beds to establish a multidimensional, coupled framework for health governance.

4 Conclusion and discussion

This study deciphers the intricate relationship between HSE and provincial health level in China through a machine learning-driven analytical framework. The findings reveal that this relationship is neither linear nor additive; instead, it is characterized by complex nonlinearities, threshold effects, and synergistic interactions across multiple environmental dimensions.

A key insight of the analysis is the prominent role of six core indicators—NMIB, UR, MPPR, RAPC, PD, and UGPR—in predicting HLI. Contrary to conventional assumptions, a higher NMIB, typically viewed as a sign of enhanced healthcare capacity, does not necessarily correlate with better health. SHAP and PDP analyses indicate a dual implication of NMIB: it reflects both healthcare supply and potential system inefficiencies or underlying health burdens. Excessive bed supply may suggest either increased disease prevalence or misallocation of medical resources.

The LOWESS curves further reveal distinct nonlinear patterns and threshold effects. For instance, UR begins to exert a positive influence on health only after surpassing the threshold range. PD exhibits two critical points—517.38 (95% CI: 472.23–609.68) and approximately 1,250 persons/km^2—indicating that moderate population agglomeration is associated with higher health levels, while excessively high densities are linked with diminishing or even negative health outcomes. Similarly, MPPR and UGPR only demonstrate significant positive associations with health level once high threshold levels are exceeded.

Interaction analysis based on PDP plots underscores that health improvements are not strongly associated with isolated environmental factors, but rather with the coordinated optimization of multiple HSE dimensions. The joint presence of high UR, high MPPR, and high RAPC is associated with higher health levels. Synergistic effects between urbanization and digital infrastructure, as well as between urbanization and physical infrastructure, are linked with higher health levels. These findings highlight the importance of integrated, system-level health governance that accounts for the interplay among various environmental components.

However, several limitations warrant attention. Although this study employs panel data (31 provinces over 11 years), the XGBoost model treats the data as pooled cross-sections, thus overlooking temporal dependencies and potential lagged effects. The model does not capture province-specific fixed effects, which may lead to biased estimations due to unobserved heterogeneity. While modeling temporal dynamics with tree-based methods remains challenging, future research could incorporate year dummies, lagged variables, or hybrid modeling techniques to better capture dynamic relationships.

Additionally, SHAP values provide transparent feature attribution but reflect statistical associations rather than causal mechanisms. As such, findings remain susceptible to unmeasured confounding. Readers should interpret the identified relationships—especially those involving complex variables such as NMIB, UR, MPPR, and RAPC—as correlational patterns, not definitive causal pathways.

Given China's vast regional diversity, exploring the heterogeneity of HSE–health relationships across different regions (e.g., eastern vs. western, urban vs. rural) could enhance the policy relevance of the findings. While this study does not conduct a stratified regional analysis due to data and methodological constraints, future research should explicitly address regional variation to better tailor policy recommendations. This limitation should be acknowledged as an avenue for future exploration.

Finally, the use of provincial-level data, while effective in capturing macro-level trends, obscures intra-urban and individual-level variation. For example, disparities between urban neighborhoods or vulnerable population subgroups remain hidden. The HSE indicator system, though comprehensive across five dimensions, is constrained by data availability. It omits critical factors such as indoor environmental quality, housing conditions, mental health status, and subjective well-being—all of which are essential for a more holistic understanding of health.

Policy implications derived from this study offer valuable directions for improving health levels through more nuanced environmental and infrastructural planning. The findings underscore that health is shaped not by isolated environmental indicators but through complex, nonlinear interactions and threshold effects across multiple dimensions of HSE. Therefore, health-oriented policy should move beyond one-size-fits-all approaches and instead embrace integrated, system-level governance. For instance, coordinated investments in urbanization, digital infrastructure, and transportation systems can generate synergistic effects that substantially improve public health, particularly when these dimensions are jointly optimized. This has important implications for cross-sectoral planning—urban development, healthcare, technology, and transportation must be aligned to maximize health returns. Moreover, these insights can inform region-specific policy strategies. Provinces with a lower level of urbanization or digital infrastructure should prioritize foundational investments, such as expanding access to primary healthcare facilities, improving road connectivity, and enhancing digital inclusion. In contrast, highly urbanized regions—especially those approaching or exceeding population density thresholds—should focus on managing urban congestion, mitigating pollution, and optimizing the distribution of healthcare resources to avoid inefficiencies or over-concentration. It is also crucial to recognize potential trade-offs. For example, excessive expansion of urban infrastructure without proper environmental safeguards may lead to resource misallocation, ecological degradation, or increased social inequality. Policymakers must therefore weigh short-term development gains against long-term health and sustainability outcomes.

Building upon the current findings, future research should aim to address several key limitations and deepen the understanding of the HSE–health nexus. First, methodological improvements are needed to better capture the temporal dynamics and fixed effects inherent in panel data. Incorporating lagged variables, time dummies, or adopting hybrid models that integrate tree-based algorithms with panel regression techniques may enhance causal inference and temporal sensitivity. Second, future studies should strive to mitigate unmeasured confounding by expanding the scope of environmental and health indicators. This includes integrating variables such as indoor air quality, housing conditions, noise exposure, green space accessibility, and subjective well-being

measures—factors currently absent due to data constraints but essential for a more holistic health assessment.

Moreover, future research should explore regional heterogeneity by conducting stratified analyses across geographic and socioeconomic divisions (e.g., eastern vs. western provinces, urban vs. rural areas). Such analysis would provide more context-sensitive insights and support differentiated policy interventions. Multi-scale data integration—linking provincial, municipal, neighborhood, and individual-level datasets—should also be prioritized. The use of high-resolution environmental data from remote sensing, geospatial platforms, and community health surveys can significantly enhance spatial granularity and policy relevance.

Lastly, incorporating behavioral and psychosocial health dimensions—such as physical activity, dietary habits, and mental health status—will further enrich the analytical framework. These enhancements will collectively support the development of more targeted, equitable, and sustainable health and urban planning strategies in diverse settings.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.jianguoyun.com/p/DQclQNkQ59HLDRjg84MGIAA.

Author contributions

HZ: Conceptualization, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing, Formal Analysis. XP: Data curation, Software, Validation, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

We acknowledge City University of Hong Kong for providing exceptional academic resources, which ensured seamless access to essential literature.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2025.1672479/full#supplementary-material

References

1. Urban Transformation and Development: China's 75-Year Journey of Economic and Social Progress——Report Xix: Urban Development Achievements in New China (2024). Available online at: https://www.stats.gov.cn/sj/sjjd/202409/t20240923_1956628.html (Accessed 23 September, 2024)

2. Wu LY Introduction to sciences of human settlements China Architecture & Building Press Beijing, China (2001) 97–112

3. Ding YP, Shi B, Su GJ, Li QQ, Meng J, Jiang YJ, et al. Assessing suitability of human settlements in high-altitude area using a comprehensive index method: a case study of Tibet, China. *Sustainability*. (2021) 13:1485. doi: 10.3390/su13031485

4. Li WJ, Li P, Feng ZM, Xiao CW. Gis-based modeling of human settlement suitability for the belt and road regions. *Int J Environ Res Public Health*. (2022) 19:6044. doi: 10.3390/ijerph19106044

5. Wang Y, Jin C, Lu MQ, Lu YQ. Assessing the suitability of regional human settlements environment from a different preferences perspective: a case study of Zhejiang Province, China. *Habitat Int*. (2017) 70:1–12. doi: 10.1016/j.habitatint.2017.09.010

6. Chen SH, Shao CF, Yu H, Gao JL. Navigating urban human settlement sustainability: a multi-indicator assessment based on sustainable development goal 11. *J Clean Prod*. (2024):472. doi: 10.1016/j.jclepro.2024.143509

7. Lin SJ, Hou LD. Sdgs-oriented evaluation of the sustainability of rural human settlement environment in Zhejiang, China. *Heliyon*. (2023) 9:E13492. doi: 10.1016/j.heliyon.2023.e13492

8. Chen JS. Temporal-spatial assessment of the vulnerability of human settlements in urban agglomerations in China. *Environ Sci Pollut Res*. (2023) 30:3726–42. doi: 10.1007/s11356-022-22420-2

9. Chen QX, Zhang KW, Zhang GY, Zhang MY. Vulnerability assessment on human settlement environment of coastal towns with entire-array-polygon method: evidence from Ninghai, China. *Environ Dev Sustain*. (2024) 27:13191–214. doi: 10.1007/s10668-023-04419-y

10. Li FZ, Zheng W, Wang Y, Liang JH, Xie S, Guo SY, et al. Urban green space fragmentation and urbanization: a spatiotemporal perspective. *Forests*. (2019) 10:333. doi: 10.3390/f10040333

11. Li GD, Zhang X, Mirzaei PA, Zhang JH, Zhao ZS. Urban heat island effect of a typical valley city in China: responds to the global warming and rapid urbanization. *Sustain Cities Soc*. (2018) 38:736–45. doi: 10.1016/j.scs.2018.01.033

12. Wang YQ, Xian CF, Jiang YQ, Pan XL, Ouyang ZY. Anthropogenic reactive nitrogen releases and gray water footprints in urban water pollution evaluation: the case of Shenzhen City, China. *Environ Dev Sustain*. (2020) 22:6343–61. doi: 10.1007/s10668-019-00482-6

13. Zhang YN, Xiang YR, Chan LY, Chan CY, Sang XF, Wang R, et al. Procuring the regional urbanization and industrialization effect on ozone pollution in Pearl River Delta of Guangdong, China. *Atmos Environ*. (2011) 45:4898–906. doi: 10.1016/j.atmosenv.2011.06.013

14. D'Amato G, Vitale C, Lanza M, Molino A, D'Amato M. Climate change, air pollution, and allergic respiratory diseases: an update. *Curr Opin Allergy Clin Immunol*. (2016) 16:434–40. doi: 10.1097/aci.0000000000000301

15. Du Y, Ding L, Na L, Sun T, Sun X, Wang LQ, et al. Prevalence of chronic diseases and alterations of gut microbiome in people of Ningxia China during urbanization: an epidemiological survey. *Frontiers in cellular and infection. Microbiology*. (2021) 11:11. doi: 10.3389/fcimb.2021.707402

16. Huang HC, Yang HL, Deng X, Zeng P, Li Y, Zhang LN, et al. Influencing mechanisms of urban Heat Island on respiratory diseases. *Iran J Public Health*. (2019) 48:1636–46. doi: 10.18502/ijph.v48i9.3023

17. Ventimiglia I, Seedat S. Current evidence on Urbanicity and the impact of Neighbourhoods on anxiety and stress-related disorders. *Curr Opin Psychiatry*. (2019) 32:248–53. doi: 10.1097/yco.0000000000000496

18. Rajagopalan S, Vergara-Martel A, Zhong J, Khraishah H, Kosiborod M, Neeland IJ, et al. The urban environment and Cardiometabolic health. *Circulation*. (2024) 149:1298–314. doi: 10.1161/CIRCULATIONAHA.123.067461

19. Xu JX, Ma J, Tao S. Examining the nonlinear relationship between neighborhood environment and residents' health. *Cities*. (2024) 152:105213. doi: 10.1016/j.cities.2024.105213

20. Bonnell LN, Littenberg B. Nonlinear relationships among the natural environment, health, and sociodemographic characteristics across US counties. *Int J Environ Res Public Health*. (2022) 19:6898. doi: 10.3390/ijerph19116898

21. Frehlich L, Christie CD, Ronksley PE, Turin TC, Doyle-Baker P, McCormack GR. The neighbourhood built environment and health-related fitness: a narrative systematic review. *Int J Behav Nutr Phys Act*. (2022) 19:124. doi: 10.1186/s12966-022-01359-0

22. Kim Y, Lee A, Cubbin C. Effect of social environments on cardiovascular disease in the United States. *J Am Heart Assoc*. (2022) 11:e025923. doi: 10.1161/JAHA.122.025923

23. Zhang K, Brook RD, Li YF, Rajagopalan S, Kim JB. Air pollution, built environment, and early cardiovascular disease. *Circ Res*. (2023) 132:1707–24. doi: 10.1161/CIRCRESAHA.123.322002

24. Cassini A, Colzani E, Pini A, Mangen MJJ, Plass D, McDonald SA, et al. Impact of infectious diseases on population health using incidence-based disability-adjusted life years (Dalys): results from the burden of communicable diseases in Europe study, European Union and European economic area countries, 2009 to 2013. *Euro Surveill*. (2018) 23:15–34. doi: 10.2807/1560-7917.ES.2018.23.16.17-00454

25. Laranjeira E, Szrek H. Going beyond life expectancy in assessments of health systems' performance: life expectancy adjusted by perceived health status. *Int J Health Econ Manage*. (2016) 16:133–61. doi: 10.1007/s10754-015-9183-z

26. Stowell JD, Kim YM, Gao Y, Fu JS, Chang HH, Liu Y. The impact of climate change and emissions control on future ozone levels: implications for human health. *Environ Int*. (2017) 108:41–50. doi: 10.1016/j.envint.2017.08.001

27. Zheng JY, Zhang N, Shen GQ, Liang FC, Zhao Y, He XC, et al. Spatiotemporal and seasonal trends of class a and B notifiable infectious diseases in China: retrospective analysis. *JMIR Public Health Surveill*. (2023) 9:9. doi: 10.2196/42820

28. Wei ZQ, Wei KK, Li Y, Nie LJ, Zhou YZ. Measurement of China's public health level: compilation and research of an index. *BMC Public Health*. (2024) 24:686. doi: 10.1186/s12889-024-18212-7

29. Cheng GS, Dou HJ, Xu S, Dai RL, Liang X, Huang YH, et al. Rural human settlement environment improvement: process, status and China's sample. *Environ Dev Sustain*. (2024) 27:17805–32. doi: 10.1007/s10668-024-04686-3

30. Hu QY, Wang C. Quality evaluation and division of regional types of rural human settlements in China. *Habitat Int*. (2020) 105:102278. doi: 10.1016/j.habitatint.2020.102278

31. Xue QR, Yang XH. Evaluation of the suitability of human settlement environment in Shanghai city based on fuzzy cluster analysis. *Therm Sci*. (2020) 24:2543–51. doi: 10.2298/TSCI2004543X

32. Oeppen J, Vaupel JW. Demography - broken limits to life expectancy. *Science*. (2002) 296:1029–31. doi: 10.1126/science.1069675

33. White KM. Longevity advances in high-income countries, 1955-96. *Popul Dev Rev*. (2002) 28:59–76. doi: 10.1111/j.1728-4457.2002.00059.x

34. Xue XH, Yang XG. Seismic liquefaction potential assessed by fuzzy comprehensive evaluation method. *Nat Hazards*. (2014) 71:2101–12. doi: 10.1007/s11069-013-0997-z

35. Wu F, Su X, Ock YS, Wang ZY. Personal credit risk evaluation model of P2p online lending based on Ahp. *Symmetry*. (2021) 13:83. doi: 10.3390/sym13010083

36. Han F, Alkhawaji RN, Shafieezadeh MM. Evaluating sustainable water management strategies using topsis and fuzzy topsis methods. *Appl Water Sci*. (2025) 15:4. doi: 10.1007/s13201-024-02336-7

37. Allee KD, Do C, Raymundo FG. Principal component analysis and factor analysis in accounting research. *J Financ Reporting*. (2022) 7:1–39. doi: 10.2308/JFR-2021-005

38. Chen CH. A novel multi-criteria decision-making model for building material supplier selection based on entropy-Ahp weighted Topsis. *Entropy*. (2020) 22:259. doi: 10.3390/e22020259

39. Hwang C-L, Yoon K. Methods for multiple attribute decision making In: C-L Hwang and K Yoon, editors. Multiple attribute decision making: Methods and applications a state-of-the-art survey. Berlin, Heidelberg: Springer Berlin Heidelberg (1981). 58–191.

40. Bentéjac C, Csörgo A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev*. (2021) 54:1937–67. doi: 10.1007/s10462-020-09896-5

41. Putatunda S, Rama K, Acm A. A Comparative Analysis of Hyperopt as against Other Approaches for Hyper-Parameter Optimization of Xgboost. 2018 International conference on signal processing and machine learning (SPML 2018) (2018). 6–10.

42. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in neural information processing systems 30 (NIPS 2017) (2017).

43. Déziel JD. Emergency medical services demand: an analysis of county-level social determinants. *Disaster Med Public Health Prep*. (2022) 17:e119. doi: 10.1017/dmp.2022.26

44. Fries JF, Koop CE, Sokolov J, Beadle CE, Wright D. Beyond health promotion: reducing need and demand for medical care. *Health Aff*. (1998) 17:70–84. doi: 10.1377/hlthaff.17.2.70

45. Zhang H, Shi LY, Yang JY, Sun G. Efficiency and equity of bed utilization in China's health institutions: based on the rank-sum ratio method. *Int J Equity Health*. (2023) 22:177. doi: 10.1186/s12939-023-01986-4

46. Ngounou BA, Oumbe HT, Fowagap JMG, Domguia EN. Is rapid urbanisation in Africa jeopardising the health and education of the population? *Rev Dev Econ*. (2025) 29:499–547. doi: 10.1111/rode.13137

47. Zhang JW, Feng XQ, Shi WH, Cui J, Peng J, Lei L, et al. Health promoting green infrastructure associated with green space visitation. *Urban For Urban Green*. (2021) 64:127237.doi: 10.1016/j.ufug.2021.127237

48. Jennings L, Omoni A, Akerele A, Ibrahim Y, Ekanem E. Disparities in Mobile phone access and maternal health service utilization in Nigeria: a population-based survey. *Int J Med Inform*. (2015) 84:341–8. doi: 10.1016/j.ijmedinf.2015.01.016

49. Sergi D, Sari IU. Prioritization of public services for digitalization using fuzzy Z-Ahp and fuzzy Z-Waspas. *Complex Intell Syst*. (2021) 7:841–56. doi: 10.1007/s40747-020-00239-z

50. Sadagheyani HE, Tatari F. Investigating the role of social media on mental health. *Ment Health Soc Inclus*. (2021) 25:41–51. doi: 10.1108/MHSI-06-2020-0039

51. Syed ST, Gerber BS, Sharp LK. Traveling towards disease: transportation barriers to health care access. *J Community Health*. (2013) 38:976–93. doi: 10.1007/s10900-013-9681-1

52. Anenberg SC, Miller J, Henze DK, Minjares R, Achakulwisut P. The global burden of transportation tailpipe emissions on air pollution-related mortality in 2010 and 2015. *Environ Res Lett*. (2019) 14:094012. doi: 10.1088/1748-9326/ab35fc

53. Konkor I. Examining the relationship between transportation mode and the experience of road traffic accident in the upper west region of Ghana. *Case Stud Transp Policy*. (2021) 9:715–22. doi: 10.1016/j.cstp.2021.03.009

54. Sohrabi S, Khreis H. Burden of disease from transportation noise and motor vehicle crashes: analysis of data from Houston, Texas. *Environ Int*. (2020) 136:136. doi: 10.1016/j.envint.2020.105520

55. Jacob N, Munford L, Rice N, Roberts J. Does commuting mode choice impact health? *Health Econ*. (2021) 30:207–30. doi: 10.1002/hec.4184

56. Shi Y, Yang JY, Shen PY. Revealing the correlation between population density and the spatial distribution of urban public service facilities with mobile phone data. *ISPRS Int J Geo Inf*. (2020) 9:38. doi: 10.3390/ijgi9010038

57. Cain DN, Mirzayi C, Rendina HJ, Ventuneac A, Grov C, Parsons JT. Mediating effects of social support and internalized Homonegativity on the association between population density and mental health among gay and bisexual men. *LGBT Health*. (2017) 4:352–9. doi: 10.1089/lgbt.2017.0002

58. Borck R, Schrauth P. Population density and urban air quality. *Reg Sci Urban Econ*. (2021) 86:103596. doi: 10.1016/j.regsciurbeco.2020.103596

59. Alhajeri NS, Dannoun M, Alrashed A, Aly AZ. Environmental and economic impacts of increased utilization of natural gas in the electric power generation sector: evaluating the benefits and trade-offs of fuel switching. *J Nat Gas Sci Eng*. (2019) 71:102969. doi: 10.1016/j.jngse.2019.102969

60. Zhao N, Li BW, Li H, Ahmad R, Peng K, Chen DY, et al. Field-based measurements of natural gas burning in Domestic Wall-mounted gas stove and estimates of climate, health and economic benefits in rural Baoding and Langfang regions of northern China. *Atmos Environ*. (2020) 229:117454. doi: 10.1016/j.atmosenv.2020.117454

61. Sun J, Luo HY. Evaluation on equality and efficiency of health resources allocation and health services utilization in China. *Int J Equity Health*. (2017) 16:127. doi: 10.1186/s12939-017-0614-y

62. Hou B, Nazroo J, Banks J, Marshall A. Are cities good for health? A study of the impacts of planned urbanization in China. *Int J Epidemiol*. (2019) 48:1083–90. doi: 10.1093/ije/dyz031

63. Tripathi S, Maiti M. Does urbanization improve health outcomes: a cross country level analysis. *Asia Pac J Reg Sci*. (2023) 7:277–316. doi: 10.1007/s41685-022-00268-1

64. Ghaedrahmati S, Alian M. Health risk assessment of relationship between air pollutants' density and population density in Tehran, Iran. *Hum Ecol Risk Assess*. (2019) 25:1853–69. doi: 10.1080/10807039.2018.1475217

65. Beenackers MA, Groeniger JO, Kamphuis CBM, Van Lenthe FJ. Urban population density and mortality in a compact Dutch city: 23-year follow-up of the Dutch globe study. *Health Place*. (2018) 53:79–85. doi: 10.1016/j.healthplace.2018.06.010

66. Niu XY, Yue YF, Zhou XG, Zhang XH. How urban factors affect the spatiotemporal distribution of infectious diseases in addition to intercity population movement in China. *ISPRS Int J Geo Inf*. (2020) 9:615. doi: 10.3390/ijgi9110615