

Article

Can Digital Infrastructure Predict Regional Innovation Capacity? Evidence Based on Machine Learning and the SHAP Explanatory Framework

Shasha Xie ¹, Shusheng Xu ¹, Wei Xu ², Guo Yu ^{1,*}, Yunli Li ¹, Jianqiu Wu ³, Minghua Xiao ⁴, Xuesong Cheng ⁵ and Peng Zhang ⁶

¹ School of Civil Engineering and Architecture, Wuhan Institute of Technology, Wuhan 430205, China; xss@wit.edu.cn (S.X.); xushusheng0926@163.com (S.X.); liyunli@wit.edu.cn (Y.L.)

² Meizhou Geological Survey Center of Guangdong Geological Bureau, Meizhou 514089, China; wallow_somnus@hotmail.com

³ China Construction Eighth Engineering Bureau Co., Ltd., Shanghai 200120, China; jianqiuw1987@163.com

⁴ China Railway 11th Engineering Bureau Group Co., Ltd., Wuhan 430061, China; cqxmh@163.com

⁵ School of Civil Engineering, Tianjin University, Tianjin 300072, China; cheng_xuesong@163.com

⁶ China Railway Design Group Co., Ltd., Tianjin 300308, China; zippor2005@sohu.com

* Correspondence: yuguo@wit.edu.cn

Abstract

Digital infrastructure is a critical foundation for promoting regional innovation and achieving sustainable development. However, existing studies have primarily focused on its impact effects, while paying limited attention to whether digital infrastructure can effectively identify future changes in regional innovation levels. Using panel data from 30 Chinese provinces during 2011–2023, this study constructs a comprehensive digital infrastructure index (*Digital*) and integrates the XGBoost and SHAP methods to evaluate the predictive capability of digital infrastructure for regional innovation. The results indicate that digital infrastructure improves the prediction accuracy of regional innovation and maintains high importance across different models and robustness tests. Further analysis reveals significant nonlinear and heterogeneous characteristics in its predictive contribution, with stronger effects observed in central and western regions, regions with relatively low levels of digital infrastructure development, and regions with higher R&D investment. These findings suggest that strengthening the construction and application of digital infrastructure can enhance regional innovation governance capacity, providing valuable insights for promoting coordinated regional innovation and sustainable development.

Keywords: digital infrastructure; regional innovation; machine learning; SHAP; predictive analysis; sustainable development



Academic Editor: Manuela Almeida

Received: 5 June 2026

Revised: 22 June 2026

Accepted: 29 June 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

1. Introduction

With the rapid development of the digital economy, digital infrastructure has gradually emerged as a critical foundation for promoting economic and social transformation [1,2]. It serves not only as a strategic prerequisite for countries seeking to strengthen their competitiveness and leadership in the digital era, but also as a key indicator of China's efforts to build a cyber power and advance the Digital China initiative. Against this backdrop, digital infrastructure—including the Internet, mobile communication networks, and data centers—has undergone continuous expansion and improvement. China has accelerated its digital transformation strategy, fostering a profound reshaping of regional

economic operations and the organization of innovation activities. At the same time, regional innovation capability has become a crucial driver of high-quality development and an important indicator for assessing regional competitiveness. Despite China's remarkable progress in innovation-driven development, substantial disparities persist across regions in terms of innovation resources, R&D investment, and levels of digitalization. Given these pronounced differences in regional innovation performance, accurately identifying the key determinants of regional innovation capability and enhancing the ability to predict future innovation levels have become important concerns for both academic researchers and policymakers [3,4].

Existing studies generally suggest a strong relationship between digital infrastructure and regional innovation performance. Specifically, (a) the development of digital infrastructure has been found to exert a significant positive effect on regional innovation [1,5], and (b) disparities in digital infrastructure can exacerbate innovation gaps across regions, while the digital divide may eventually be transformed into a digital dividend once digital infrastructure development surpasses a certain threshold [5]. Consequently, digital infrastructure has become a key topic in the literature on regional innovation. However, most existing studies have relied on traditional analytical approaches, such as panel regression models [1] and spatial econometric methods [6], to examine the impact of digital infrastructure on regional innovation. These approaches primarily focus on identifying causal relationships among variables. In contrast, limited attention has been paid to whether digital infrastructure possesses strong predictive power for future regional innovation performance or whether complex nonlinear relationships exist between digital infrastructure and regional innovation.

Meanwhile, with the rapid advancement of artificial intelligence and big data technologies, machine learning techniques have been increasingly adopted in forecasting research [7]. Compared with traditional regression models, machine learning approaches offer several advantages, including their data-driven nature, superior capability in handling high-dimensional data, and enhanced ability to capture complex interactions and nonlinear patterns. Among these methods, the XGBoost algorithm has demonstrated outstanding predictive performance and robustness, leading to its widespread application in fields such as economics, finance, and energy management. Nevertheless, its application to predicting regional innovation performance based on digital infrastructure remains relatively limited. Therefore, it is necessary to integrate machine learning techniques with interpretable analytical methods to investigate the relationship between digital infrastructure and regional innovation from a predictive perspective.

Against this backdrop, a growing body of literature has examined the relationship between digital infrastructure and regional innovation. Existing studies can be broadly categorized into three strands: the measurement of digital infrastructure, the relationship between digital infrastructure and regional innovation, and the application of machine learning in economic forecasting:

(a) Measurement and Economic Effects of Digital Infrastructure.

There is no consensus regarding the measurement of digital infrastructure. For example, Dan C. et al. [4] employed indicators such as Internet usage, Internet penetration rates, and the number of data centers to investigate the relationship between digital infrastructure and regional innovation. Similarly, Zhai C. [8] and Nie C. et al. [5] used policy initiatives such as "Broadband China" and "Smart City" programs as proxies for digital infrastructure development. Chao X. et al. [9] utilized Python (version 3.9)-based text mining techniques to analyze the frequency of keywords such as "5G" and "data centers" in government work reports. Other scholars have measured digital infrastructure development through either a single indicator [10] or a composite index comprising multiple indicators [11]. Regarding its

economic effects, existing research has primarily focused on the role of digital infrastructure in enhancing green total factor productivity, narrowing the urban–rural income gap, and influencing carbon emissions. Wang Q. et al. [12] empirically examined the impact and underlying mechanisms of digital infrastructure on green total factor productivity and found that digital infrastructure indirectly promotes productivity growth by stimulating technological innovation. Liang K. et al. [13] demonstrated that digital infrastructure contributes to reducing the urban–rural income gap through promoting non-agricultural employment, enhancing technological innovation, and facilitating industrial upgrading. Using national-level panel data, Che S. et al. [14] investigated the impact of digital infrastructure on carbon emissions and its transmission mechanisms, concluding that digital infrastructure increases global carbon emissions through capital accumulation and fossil energy consumption channels.

(b) The Relationship Between Digital Infrastructure and Regional Innovation.

The relationship between digital infrastructure and regional innovation has attracted extensive scholarly attention. Existing studies generally suggest that digital infrastructure promotes regional innovation by enhancing knowledge creation, improving labor productivity, reducing the costs of knowledge acquisition, facilitating green industrial transformation, and fostering cross-regional collaborative innovation among firms. For instance, Kafourous et al. [15] argued that the Internet enhances firms' innovative capabilities by improving efficiency across three key dimensions: cost, time, and quality through information search and communication functions. Using the “Broadband China” policy as a quasi-natural experiment, Zhao X. et al. [16] found that digital infrastructure enhances labor productivity by strengthening urban innovation capacity. Employing county-level data from the United States, Xu [17] demonstrated that Internet access reduces information acquisition and dissemination costs, thereby stimulating innovation activities. Guo B. et al. [18] examined the impact of digital infrastructure on corporate green transformation among Chinese A-share listed industrial firms using a time-varying difference-in-differences approach. Their findings indicate that digital infrastructure promotes green transformation by facilitating green technological innovation and alleviating financing constraints. Similarly, Tian X. et al. [19] highlighted that investment in digital infrastructure improves information transmission efficiency and thereby enhances innovation cooperation among firms across different regions.

In addition to examining the direct effects of digital infrastructure on innovation, some scholars have explored the heterogeneity of this relationship from multiple perspectives. Existing evidence suggests that the impact of digital infrastructure on regional innovation varies significantly across regions. Differences in economic development levels [7], market integration [7,19], city size [20], and industrial structure [7,21] lead to heterogeneous innovation outcomes associated with digital infrastructure development.

(c) Applications of Machine Learning in Economic Forecasting.

With the rapid advancement of artificial intelligence technologies, machine learning (ML) has emerged as an important methodological tool in economic forecasting. Compared with traditional econometric models, ML approaches can automatically learn complex nonlinear relationships without imposing restrictive model assumptions and possess strong capabilities in feature extraction and prediction. Consequently, they have been increasingly applied in both causal inference and forecasting research. For example, Zhang H. et al. [22] employed a Double Machine Learning (DML) framework to investigate the effects of digital infrastructure on the coordinated governance of pollution reduction and carbon mitigation, thereby advancing research on high-dimensional policy evaluation and causal inference. Similarly, Fang H. et al. [23] and Huang J. et al. [24] adopted DML models to

evaluate the impacts of digital infrastructure on urban green innovation and energy-use efficiency, respectively.

Beyond causal analysis, machine learning techniques have been widely utilized in areas such as energy price forecasting, risk management, and macroeconomic prediction [25,26]. For instance, Ji T. et al. [27] applied deep learning and machine learning models to forecast energy demand and CO₂ emissions in China's transportation sector using variables such as population, GDP per capita, and vehicle kilometers traveled. Their results indicate that transportation-related energy consumption and carbon emissions could increase by approximately 3.5 times by 2050, demonstrating the effectiveness of machine learning methods in energy-economic forecasting. Likewise, Xie H. et al. [28] employed machine learning techniques to forecast China's macroeconomic performance, using quarterly real GDP growth as the prediction target and comparing the forecasting accuracy of various machine learning models with traditional econometric approaches. The findings reveal that machine learning models achieve superior predictive accuracy during periods of economic stability, with average forecast errors generally lower than those of conventional models. Furthermore, the study utilized explainable machine learning techniques to identify the key drivers of GDP growth, providing new insights into the analysis of macroeconomic fluctuations.

In summary, although the existing literature has generated substantial insights into the relationship between digital infrastructure and regional innovation, several important gaps remain. First, most previous studies have employed traditional econometric approaches, such as panel regression models, difference-in-differences (DID) methods, and spatial econometric models, to investigate the impact of digital infrastructure on regional innovation and its underlying mechanisms. These studies primarily focus on causal identification and statistical significance, while paying limited attention to whether digital infrastructure possesses predictive power for regional innovation or can serve as a key predictor of future changes in innovation performance. Second, although machine learning techniques have been widely applied in economic forecasting and have demonstrated strong predictive performance in areas such as energy demand forecasting, economic growth prediction, and risk management, their application to research on digital infrastructure and regional innovation remains relatively limited. In particular, there is a lack of studies that examine the importance of digital infrastructure from a predictive perspective. Third, existing studies mainly concentrate on whether digital infrastructure affects regional innovation, while insufficient attention has been devoted to its relative predictive importance among a broad set of socioeconomic determinants. It remains unclear whether digital infrastructure constitutes a primary or secondary driver of regional innovation and how its predictive contribution compares with that of traditional innovation determinants, such as R&D investment, human capital, and financial development. Moreover, digital infrastructure is inherently multidimensional. The predictive effects of infrastructure-related indicators and application-related indicators may differ substantially. However, most prior studies have relied on composite indices or single indicators, with limited efforts devoted to identifying the predictive contribution and heterogeneous effects of individual dimensions and sub-indicators.

To address these gaps, this study focuses on the predictive relationship between digital infrastructure and regional innovation capability and seeks to answer the following research questions:

- (a) Does digital infrastructure contribute to predicting regional innovation capability? How does its predictive importance compare with that of other socioeconomic characteristics?
- (b) How do different machine learning models, including XGBoost, Random Forest, and LightGBM, perform in predicting regional innovation capability? Which model achieves the highest predictive accuracy?

- (c) Does the predictive effect of digital infrastructure on regional innovation exhibit nonlinear characteristics? Which specific indicators related to digital infrastructure construction and digital application possess stronger predictive power?
- (d) Does the predictive importance of digital infrastructure vary across regions and stages of development? Specifically, does its predictive contribution differ significantly across regions with different levels of digital infrastructure development and R&D investment?

To answer these questions, this study employs panel data from 30 provincial-level regions in China covering the period from 2011 to 2023. A comprehensive digital infrastructure index is constructed, and three machine learning models—OLS, XGBoost, Random Forest, and LightGBM—are utilized to predict regional innovation capability. Furthermore, the SHAP (Shapley Additive Explanations) framework is applied to identify key predictive factors and uncover potential nonlinear relationships. By integrating machine learning and explainable artificial intelligence techniques, this study aims to provide a novel perspective and empirical evidence for understanding the relationship between digital infrastructure and regional innovation.

2. Methods and Data

2.1. Research Framework

The research framework of this study consists of four main steps. First, using panel data from 30 Chinese provinces covering the period 2011–2023, an indicator system is constructed from two dimensions—digital infrastructure construction and digital infrastructure application. The entropy weighting method is then employed to calculate a comprehensive digital infrastructure index. Based on this framework, a regional innovation forecasting dataset is developed, in which regional innovation performance serves as the prediction target, while digital infrastructure indicators and related socioeconomic variables are used as predictive features.

Second, machine learning models, including XGBoost, Random Forest, and LightGBM, are employed to predict regional innovation performance. Model performance is evaluated using multiple metrics, including the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

Finally, the SHAP (Shapley Additive Explanations) method is applied to identify key predictive factors and quantify their contributions to regional innovation prediction. In addition, robustness tests and heterogeneity analyses are conducted to verify the reliability and generalizability of the empirical findings.

2.2. Sample Selection and Data Sources

Due to the approximately two-year publication lag of official statistical yearbooks, the most recent comprehensive and authoritative annual data currently available extend to 2023. Accordingly, this study employs panel data from 30 provincial-level regions in China (excluding Tibet, Hong Kong, Macao, and Taiwan) covering the period from 2011 to 2023. A total of 390 valid observations are obtained. The data are collected from the China Energy Statistical Yearbook, the China Statistical Yearbook, and the statistical yearbooks of individual provinces and municipalities. A small number of missing observations (Some predictor variables contain missing values for certain provinces in 2022 or 2023. Across all observed variables, a total of 21 values were missing, accounting for 0.32% of all observations. The variables requiring interpolation were *WEB*, *PC*, and *PSI*.) are addressed using linear interpolation for continuous variables. To evaluate the predictive performance of the machine learning models, the dataset is divided chronologically. Observations from 2011 to 2021 are used as the training set, data for 2022 are used as the validation set for

model tuning and selection, and data for 2023 are reserved as the test set for out-of-sample evaluation. Finally, all variables are standardized to eliminate scale differences and improve model performance.

2.3. Variable Definition and Measurement

2.3.1. Regional Innovation Capability

Regional Innovation Capability ($\ln INN$). Patents are widely recognized as one of the most important outputs of innovation activities and are among the most commonly used and representative indicators for measuring regional innovation capability. Compared with patent applications, granted patents undergo substantive examination procedures and therefore provide a more accurate reflection of the quality of innovation outcomes and their practical transformation into productive results. This approach also helps avoid the overestimation of innovation output caused by strategic patent applications, duplicate filings, or other non-substantive applications. Accordingly, this study uses the number of granted patents in each region as a proxy for regional innovation capability. To mitigate potential issues related to data volatility and heteroscedasticity, the variable is transformed using the natural logarithm, denoted as $\ln INN$.

Furthermore, given that this study aims to develop a machine learning-based framework for predicting regional innovation capability, the dependent variable should possess characteristics such as observability, quantifiability, and interpretability. The number of granted patents is derived from reliable and standardized statistical sources and directly reflects the level of regional innovation output. As such, it not only satisfies the requirements of machine learning model training and prediction but also facilitates subsequent SHAP-based interpretability analysis for identifying key predictive factors. Therefore, employing the number of granted patents as a measure of regional innovation capability is both theoretically justified and practically appropriate.

2.3.2. Digital Infrastructure

Digital Infrastructure. Considering data availability and drawing on the framework proposed by He X. et al. [8,27], this study measures digital infrastructure from two dimensions: infrastructure construction and infrastructure application (Table 1).

Table 1. Digital Infrastructure Indicator System.

Dimension	Variable	Measurement Method	Unit
Infrastructure Construction	<i>LDOCD</i>	Length of long-distance optical cables/Administrative land area	km/km ²
	<i>BAR</i>	Number of broadband Internet subscribers/Total population	subscribers per capita
	<i>MPEC</i>	Mobile telephone exchange capacity/Total population	lines per capita
	<i>PC</i>	Total number of computers/Total population	units per capita
Infrastructure Application	<i>DOM</i>	Total number of domain names/Total population	domains per capita
	<i>WEB</i>	Total number of webpages/Total population	webpages per capita
	<i>MPPR</i>	Number of mobile phone users/Total population	%
	<i>IPR</i>	Number of Internet users/Total population	%

Notes: Italic fonts are used to denote all research variables in this paper.

Specifically, the digital infrastructure indicator system comprises eight indicators: long-distance optical cable density (*LDOCD*), broadband access rate (*BAR*), mobile telephone exchange capacity (*MPEC*), number of computers per capita (*PC*), number of domain names (*DOM*), number of webpages (*WEB*), mobile phone penetration rate (*MPPR*), and Internet penetration rate (*IPR*). Detailed definitions, calculation methods, and variable descriptions are presented in Table 1.

To comprehensively capture the level of digital infrastructure development, the entropy weighting method is employed to calculate a composite digital infrastructure index (*Digital*). This approach objectively determines indicator weights based on the information contained in the data and reduces subjective bias in index construction. The resulting composite index is subsequently used to examine its relationship with regional innovation capability and to evaluate its predictive importance within the machine learning framework.

2.3.3. Other Predictive Variables

To improve the predictive accuracy of the models, this study incorporates eight additional explanatory variables that have been widely recognized in the literature as important determinants of regional innovation performance.

Urbanization Rate (*URB*) is measured as the proportion of the urban population to the total population. The urbanization process facilitates knowledge agglomeration and factor mobility, thereby promoting regional innovation. R&D Intensity (*RD*) is measured by the ratio of R&D expenditure to GDP and is widely regarded as a key driver of innovation performance. Industrial Structure (*PSI*) is proxied by the share of value added generated by the tertiary sector in GDP. Government Intervention (*GI*), This indicator is defined as the ratio of local government general budget expenditure to regional GDP. Financial Development (*FD*) is measured by the ratio of the combined balance of deposits and loans of financial institutions to GDP. Financial support provides essential resources for innovation activities and technological advancement. Level of Openness (*LOU*) is measured as the ratio of total regional imports and exports to GDP, reflecting the extent to which a region is integrated into international markets and exposed to external knowledge spillovers. Human Capital (*HC*) is measured by the ratio of university students enrolled in higher education institutions to the total population. As a critical source of knowledge creation and technological progress, human capital constitutes one of the fundamental drivers of innovation.

2.3.4. Descriptive Statistics of Variables

This section presents the descriptive statistics of all variables used in the study (Table 2). The reported statistics include the mean, standard deviation, minimum value, and maximum value, which provide insights into the distributional characteristics and variability of the data.

Regional Innovation Capability. Before logarithmic transformation, the mean value of *INN* is 78,777.21, with a standard deviation of 120,695.25, indicating substantial disparities in innovation performance across regions. After logarithmic transformation, the mean of $\ln(INN)$ is 10.367 and the standard deviation decreases to 1.481, suggesting that the transformation effectively mitigates data skewness and reduces heteroscedasticity.

Digital Infrastructure. The mean values of *Digital*, *LDOCD*, and *BAR* are 0.129, 0.194, and 0.527, respectively. Their standard deviations remain within a reasonable range, indicating the existence of regional heterogeneity in digital infrastructure development. In contrast, the *WEB* variable exhibits a mean value of 241.606 and a notably high standard deviation of 792.653, reflecting substantial disparities in the distribution of online resources across regions.

Other Predictive Variables. Variables such as *URB*, *RD*, and *HC* display relatively small standard deviations, suggesting a comparatively stable distribution among provinces. Overall, the values of all variables fall within reasonable ranges, and no extreme outliers are observed in the sample.

Table 2. Descriptive Statistics.

Variable Category	Variable	Mean	Standard Deviation	Minimum	Maximum
Regional Innovation Capability	<i>INN</i>	78,777.205	120,695.248	447.000	872,219.000
	<i>ln(INN)</i>	10.367	1.481	6.102	13.679
Digital Infrastructure	<i>Digital</i>	0.129	0.113	0.010	0.895
	<i>LDOCD</i>	0.194	0.133	0.022	0.889
	<i>BAR</i>	0.527	0.250	0.096	1.112
	<i>MPEC</i>	1.780	0.586	0.806	4.922
	<i>DOM</i>	0.023	0.042	0.001	0.338
	<i>WEB</i>	241.606	792.653	0.017	6445.589
	<i>MPPR</i>	104.568	25.319	52.040	188.620
	<i>IPR</i>	0.258	0.124	0.052	0.558
	<i>PC</i>	28.028	12.644	6.636	81.000
Other Predictive Variables	<i>URB</i>	0.601	0.121	0.344	0.942
	<i>RD</i>	0.018	0.011	0.004	0.063
	<i>PSI</i>	0.398	0.102	0.026	0.613
	<i>GI</i>	0.252	0.108	0.105	0.749
	<i>FD</i>	3.405	1.030	1.694	7.545
	<i>LOU</i>	0.263	0.267	0.007	1.389
	<i>HC</i>	0.021	0.006	0.008	0.044

2.4. Research Methods

2.4.1. Entropy Weight Method

To objectively measure the level of digital infrastructure development and avoid subjective bias in weight assignment, this study adopts the entropy weight method to construct a composite digital infrastructure index (*Digital*), following the approach of [5]. Based on information entropy theory, the entropy weight method assigns weights in a fully objective manner. Specifically, indicators with greater variation across samples possess stronger discriminatory power, resulting in lower entropy values and higher information redundancy, and are therefore assigned larger weights. This method effectively extracts information embedded in the original panel data and is well suited to multi-provincial, multi-year observational samples. Let the original data matrix be denoted as $X = (x_{ij})_{m \times n}$, where (m) represents the number of samples and (n) represents the number of indicators. The calculation procedure is as follows. First, all indicators are standardized using the Min–Max normalization method (Equation (1)). Second, based on the standardized data, the proportion of the i -th sample under the j -th indicator (p_{ij}) is calculated for subsequent entropy estimation (Equation (2)). Third, the information entropy of each indicator is calculated. A larger entropy value indicates a smaller amount of information provided by the corresponding indicator (Equation (3)). Fourth, the weight of each indicator is determined according to its information redundancy (Equation (4)). Finally, the composite digital infrastructure index for each province and year is calculated as the weighted sum of all indicators (Equation (5)). The weights of all indicators are presented in Appendix A Table A1.

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (1)$$

where $\max(x_j)$ and $\min(x_j)$ denote the maximum and minimum values of indicator j across all samples, respectively. After standardization, the values of all indicators are transformed into the interval $[0, 1]$.

$$p_{ij} = \frac{x'_{ij}}{\sum_{i=1}^m x'_{ij}} \quad (2)$$

where p_{ij} denotes the proportion of the i -th sample under the j -th indicator.

$$e_j = -\frac{1}{\ln(m)} \sum_{i=1}^m p_{ij} \ln(p_{ij}) \quad (3)$$

where (e_j) denotes the information entropy of indicator (j) ; (m) represents the number of samples; and (p_{ij}) denotes the proportion of the (i) -th sample under the (j) -th indicator.

$$w_j = \frac{1 - e_j}{\sum_{j=1}^n (1 - e_j)} \quad (4)$$

where (w_j) denotes the weight of indicator (j) ; $(1 - e_j)$ represents the information redundancy of the indicator (the smaller the entropy value, the greater the redundancy and the higher the corresponding weight); and (n) denotes the number of indicators.

$$Digital = \sum_{j=1}^n x'_{ij} w_j \quad (5)$$

where *Digital* denotes the comprehensive digital infrastructure index for each province in each year.

2.4.2. Prediction Models

This study constructs the following regional innovation capability prediction model:

$$\ln INN_{i,t} = f(D_{i,t}) + \varepsilon_{i,t} \quad (6)$$

where $D_{i,t}$ denotes the set of variables for province i in year t , including digital infrastructure and other predictive variables; $\ln INN_{i,t}$ represents the logarithm of regional innovation capability in province i at year t ; $f(\cdot)$ denotes the prediction model employed; and $\varepsilon_{i,t}$ represents the prediction residual.

This study employs the baseline OLS model, the Random Forest (RF) model, the XGBoost model, and the LightGBM model for comparative prediction analysis. OLS estimates a linear relationship by minimizing the sum of squared residuals between the observed and predicted values of the dependent variable, thereby quantifying the average linear effect of the independent variables on the dependent variable. The RF model integrates multiple decision trees through a voting mechanism, exhibiting strong resistance to overfitting and high stability, making it suitable for a wide range of conventional datasets. XGBoost is an advanced implementation of gradient boosting trees with enhanced regularization mechanisms, offering high predictive accuracy and strong performance in structured data classification and regression tasks. LightGBM is a lightweight gradient boosting framework characterized by fast computational speed and low memory consumption, making it particularly advantageous for large-scale data applications. These models are widely recognized as representative machine learning algorithms and are frequently combined with SHAP methods for model interpretation.

2.4.3. Explainable Machine Learning Based on SHAP

Traditional linear regression models can only capture the average linear marginal effects of explanatory variables and are unable to adequately characterize nonlinear relationships and interaction effects. In contrast, although ensemble machine learning models generally achieve higher predictive accuracy, they often suffer from the “black-box” problem, making it difficult to provide economically meaningful interpretations of individual variables at the sample level. Therefore, this study introduces the SHAP (Shapley Addi-

tive Explanations) framework to decompose and interpret the prediction mechanisms of machine learning models.

The concept of SHapley Additive exPlanations (SHAP) originates from the Shapley value theory proposed by Lloyd Shapley in 1953. It was initially developed to address the problem of fair payoff allocation in cooperative game theory. Subsequently, this theory was successfully extended to the field of machine learning (SHAP treats all features in the model as players in a cooperative game, with the model prediction for a given sample regarded as the total payoff. The SHAP value of an individual feature is equivalent to its average marginal contribution to the payoff across all possible feature coalitions.), where it has been widely used to address the “black-box” problem of machine learning models. The mathematical formulation is as follows [29]:

$$f(x) = \phi_0 + \sum_{i=1}^p \phi_i(x) \quad (7)$$

where $f(x)$ denotes the prediction result of the machine learning model for a given sample. This prediction can be decomposed into the sum of a baseline value and the SHAP values of all features.

$$\phi_*^j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [v(S \cup \{j\}) - v(S)] \quad (8)$$

where (ϕ^j) denotes the SHAP value of variable (j) , representing its average marginal contribution to the overall prediction. A larger $(|\phi^j|)$ indicates a greater contribution of the variable to the prediction target. Specifically, $(\phi^j > 0)$ indicates a positive effect of the variable on the prediction target, whereas $(\phi^j < 0)$ indicates a negative effect. (F) denotes the set of all variables, and $(|F|)$ represents the total number of variables. (F/j) denotes the set of all variables excluding variable (j) . $(v(S \cup \{j\}) - v(S))$ represents the marginal gain generated by adding variable (j) to subset (S) . The term $(\frac{|S|!(|F| - |S| - 1)!}{|F|!})$ denotes the weight assigned to each subset, which corresponds to the probability that variable (j) is added as the $((|S| + 1))$ -th element in the subset, ensuring that the contributions of all possible feature orderings are fairly weighted. Overall, the SHAP method is model-agnostic and can be applied to any machine learning model. Moreover, SHAP provides global comparability, enabling the ranking and selection of important predictive factors based on their SHAP values and facilitating the identification of key indicators for predicting regional innovation levels.

2.4.4. Evaluation of Predictive Performance

The panel dataset used in this study has a temporal structure. Random data partitioning may destroy temporal dependencies and lead to data leakage, thereby overestimating the predictive performance of the models. Therefore, the dataset is divided according to chronological order into a training set (data up to and including 2021, used for model training), a validation set (data from 2022, used for hyperparameter tuning), and a test set (used to evaluate model generalization performance).

To comprehensively evaluate the predictive performance of the models, this study employs the following evaluation metrics [29]. The coefficient of determination (R^2) measures the proportion of variation in the dependent variable explained by the model, with values closer to 1 indicating better model fit. The root mean square error (RMSE) measures the difference between predicted and actual values, with smaller values indicating higher predictive accuracy. The mean absolute percentage error (MAPE) represents the prediction error as a percentage of the actual value, with smaller values indicating better predictive performance. The mean absolute error (MAE) measures the average absolute difference

between predicted and actual values, with smaller values indicating higher prediction accuracy, and an ideal value of 0.

3. Empirical Results

3.1. Model Prediction Results

Table 3 presents a comparison of two categories of models: models incorporating all predictors and models excluding the *Digital* variable.

Table 3. Model Prediction Performance.

Category	Metric	OLS	XGBoost	RF	LightGBM
All predictor variables	R ²	0.735	0.903	0.888	0.844
	RMSE	0.620	0.374	0.406	0.476
	MAE	0.505	0.267	0.284	0.357
	MAPE	4.52%	2.53%	2.74%	3.36%
All predictor variables excluding Digital	R ²	0.714	0.890	0.857	0.823
	RMSE	0.792	0.377	0.450	0.503
	MAE	0.645	0.327	0.364	0.384
	MAPE	6.62%	2.89%	3.31%	3.60%

Note: Due to space limitations, only the performance of the XGBoost model at different stages is reported. The R² values of the XGBoost model on the training and validation sets are 0.912 and 0.899, respectively, differing from that of the test set by less than 1%. This indicates that the model exhibits consistent performance across different datasets and does not show substantial discrepancies attributable to data partitioning.

(a) All predictor variables

In terms of overall predictive performance, XGBoost achieves the best results, with an R² value of 0.903, making it the only model with an R² exceeding 0.9. This indicates that the model is capable of explaining more than 90% of the variation in the dependent variable. Moreover, XGBoost yields the lowest values of RMSE (0.374), MAE (0.267), and MAPE (2.53%) among all models, demonstrating the highest prediction accuracy. The RF model ranks second, with an R² value of 0.888. Although its error metrics are slightly higher than those of XGBoost, it still exhibits strong predictive capability. LightGBM achieves an R² of 0.844, indicating a moderate level of predictive performance. Among the three models, XGBoost delivers the best performance in predicting regional innovation capability, benefiting from its gradient boosting framework and regularization mechanism.

(b) Models excluding the Digital variable

When the Digital variable is excluded, the predictive performance of all models deteriorates. Specifically, the R² values decrease by 0.013–0.031, while RMSE, MAE, and MAPE all increase. Among the models, the RF model exhibits the most pronounced decline, with its R² decreasing from 0.888 to 0.857. These results indicate that digital infrastructure provides additional predictive information for regional innovation and contributes to improving both model explanatory power and predictive accuracy. The findings further confirm the important role of digital infrastructure in forecasting regional innovation.

3.2. Model Interpretability Analysis

3.2.1. Global Analysis of SHAP Values

If the absolute SHAP value of a given indicator ranks higher among the absolute SHAP values of all indicators, the indicator is considered more important in predicting changes in the dependent variable. Table 4 presents the SHAP values and the rankings based on absolute SHAP values of different indicators across different models.

Table 4. SHAP-Based Interpretability Analysis of All Indicators Across Different Models.

Variable	XGBoost		RF		LightGBM	
	SHAP Value	Rank	SHAP Value	Rank	SHAP Value	Rank
<i>Digital</i>	0.293	3	0.225	3	0.382	3
<i>URB</i>	0.077	7	0.027	7	0.049	8
<i>RD</i>	0.386	2	0.635	1	0.586	1
<i>PSI</i>	0.088	6	0.071	5	0.102	5
<i>GI</i>	0.476	1	0.443	2	0.436	2
<i>FD</i>	0.032	8	0.017	8	0.103	4
<i>LOU</i>	0.115	5	0.062	6	0.092	7
<i>HC</i>	0.133	4	0.108	4	0.010	6

First, *GI* and *RD* are the two most important predictors of regional innovation capability. *GI* has the highest SHAP value in the XGBoost model and ranks second in both the RF and LightGBM models. *RD* ranks first in the RF and LightGBM models and second in the XGBoost model. The SHAP values of these two variables are significantly higher than those of the other variables across all three models, indicating that they make the greatest contribution to the prediction of regional innovation capability and are key predictive indicators.

Second, *Digital* exhibits strong predictive power for regional innovation capability. The *Digital* index consistently ranks third across all three models, with SHAP values of 0.293 (XGBoost), 0.225 (RF), and 0.382 (LightGBM). This suggests that Digital is the third most important predictor used by the models and contributes substantially to prediction accuracy.

Third, the predictive importance of the variables exhibits a clear tiered structure. Taking the XGBoost model as an example, the first tier consists of *GI*, *RD*, and *Digital*, whose SHAP values are substantially higher than those of the remaining variables. The second tier includes *HC*, *PSI*, and *LOU*, which make moderate contributions to prediction. The third tier comprises *URB* and *FD*, which rank relatively low and have small SHAP values, indicating limited contributions to the prediction of regional innovation capability.

3.2.2. Analysis of the Predictive Effects of Individual Indicators

Given that this study has constructed the digital infrastructure index from both the construction and application dimensions, it is possible to identify the relative contributions of different dimensions and specific indicators to the prediction of regional innovation. Therefore, To further investigate the predictive power of individual digital infrastructure indicators on regional innovation capability and identify which types of digital infrastructure indicators possess stronger predictive effectiveness, this study constructs eight separate XGBoost prediction models using the eight original digital infrastructure indicators. At the same time, the original digital infrastructure indicators exhibit a certain degree of multicollinearity (the correlation coefficients and VIF values for the eight indicators are reported in Appendix A Table A2), each indicator is incorporated into an independent prediction model for comparative analysis. Due to space limitations, Figure 1 presents the SHAP dependence plots for the five indicators whose contributions (measured by the absolute SHAP values) rank among the top three in their respective models: *LDOCD*, *BAR*, *DOM*, *MPPR*, and *IPR*.

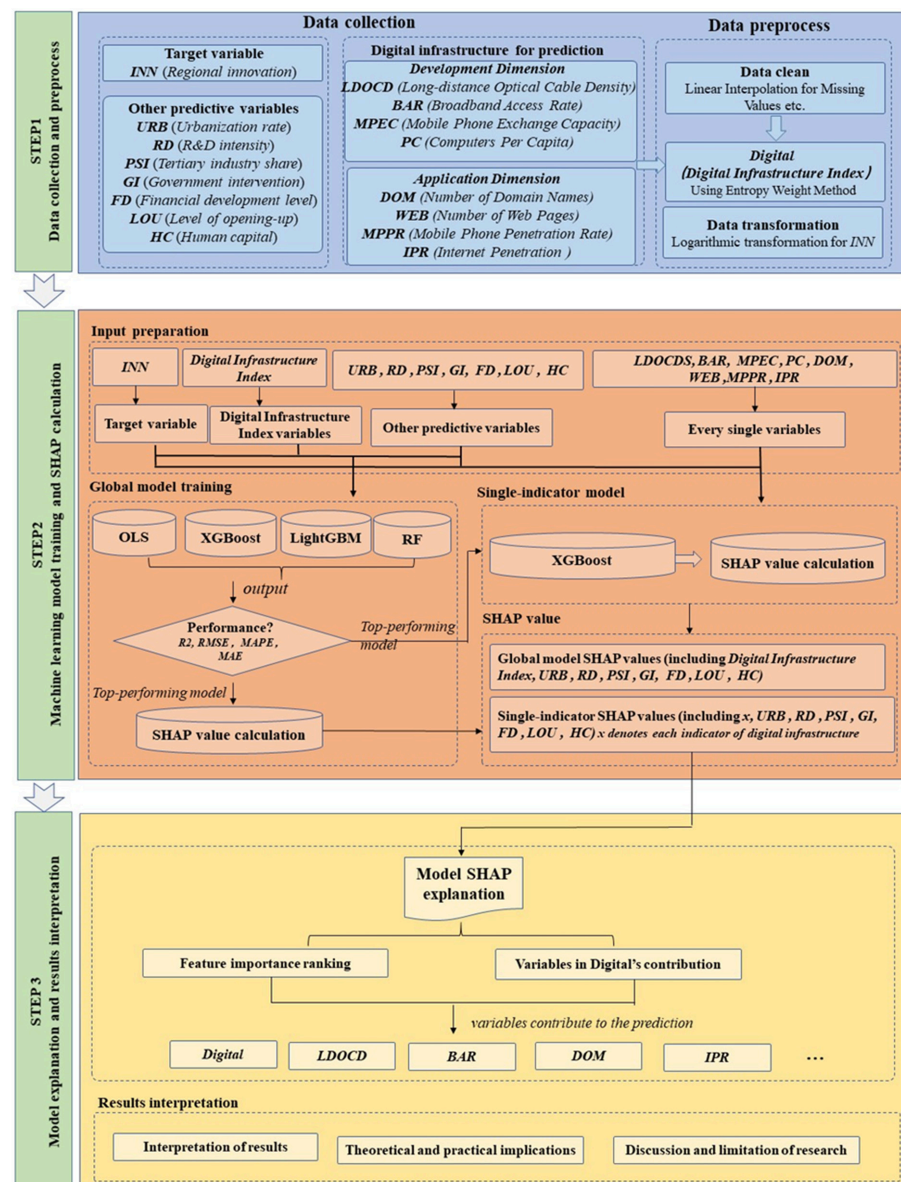


Figure 1. Overall Analytical Framework. Note: The ellipsis represents the remaining variables (WEB, MPPR, MPEC, PC) associated with Digital.

(A) Digital Infrastructure Construction. Figure 2a presents the SHAP dependence plot for *LDOCD*, while Figure 2b presents that for *BAR*. The contribution pattern of *LDOCD* within the trained model exhibits a clear nonlinear increasing trend. When *LDOCD* is at a relatively low level (below 0.19, $(0.19 + 0 \times 0.13)$), its SHAP values are mostly negative, indicating that weak digital infrastructure is associated with lower regional innovation capability. As *LDOCD* increases, its SHAP value rises rapidly and changes from negative to positive when it approaches the sample mean. Thereafter, the SHAP value continues to increase.

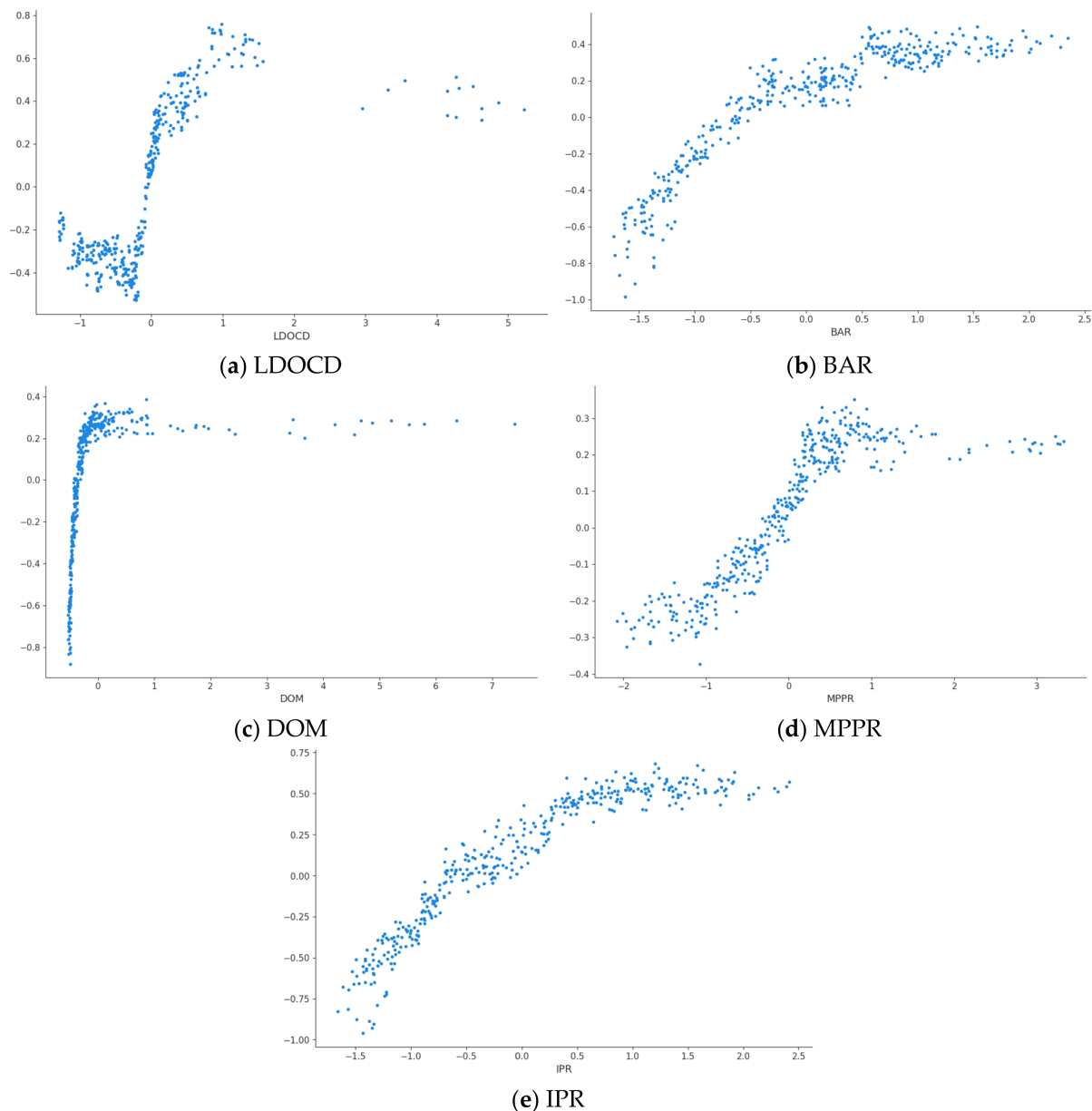


Figure 2. SHAP Dependence Plots of Individual Variables.

Similarly, the effect of *BAR* on regional innovation capability also exhibits a clear nonlinear increasing pattern. When *BAR* is at a relatively low level (below 0.40), its SHAP values are mostly negative, indicating that insufficient broadband Internet access is associated with lower regional innovation capability. As *BAR* increases, its contribution rises rapidly and turns from negative to positive when it approaches the sample mean ($0.53, (0.53 + 0 \times 0.25)$). Subsequently, the SHAP value continues to increase. However, when *BAR* exceeds a relatively high level (approximately 0.65), the upward trend gradually flattens, indicating diminishing marginal returns. This pattern suggests that the positive contribution of *BAR* to model predictions becomes less pronounced at higher values.

(B) Digital Infrastructure Application. Figure 2c presents the SHAP dependence plot for *DOM*, Figure 2d for *MPPR*, and Figure 2e for *IPR*. All three variables exhibit nonlinear contribution patterns with respect to the prediction target.

First, The relationship between *DOM* and its SHAP values exhibits a clear nonlinear increasing pattern, When *DOM* is at a relatively low level, its SHAP values are mostly negative. As *DOM* increases, its SHAP value rises rapidly and turns from negative to

positive when it approaches the sample mean ($0.023, (0.023 + 0 \times 0.042)$). Thereafter, the SHAP value fluctuates within a relatively high range. When *DOM* exceeds a relatively high level (approximately 0.63), its SHAP value tends to stabilize and no longer exhibits sustained rapid growth.

Second, the relationship between *MPPR* and its SHAP values also exhibits a nonlinear increasing pattern. When *MPPR* is at a relatively low level, its SHAP values are mostly negative, indicating that a low level of mobile Internet penetration is associated with lower regional innovation capability. As *MPPR* increases, its contribution turns from negative to positive and continues to rise, suggesting that higher *MPPR* values are associated with more positive SHAP contributions within the model. When *MPPR* reaches a relatively high level (approximately 104.57, $(104.57 + 0 \times 25.32)$), the growth of the SHAP value gradually slows down, indicating that the upward trend in SHAP values gradually flattens.

Finally, the relationship between *IPR* and its SHAP values exhibits a strong nonlinear positive pattern. When *IPR* is at a relatively low level, its SHAP values are mostly negative, indicating that insufficient Internet penetration is associated with lower regional innovation capability. As *IPR* increases, its SHAP value rapidly changes from negative to positive and continues to rise, with a larger increase than that observed for the other application-related indicators. When *IPR* reaches a relatively high level (approximately 0.38, $(0.26 + 1 \times 0.12)$), the SHAP value remains stable within a relatively high range, indicating persistently positive SHAP contributions within the trained model.

In terms of contribution magnitude, the SHAP values of *IPR* are generally higher than those of *DOM* and *MPPR*, suggesting that *IPR* plays the most important role in explaining the model's prediction results among the indicators in the digital infrastructure application dimension.

3.3. Robustness Tests

3.3.1. Model Replacement

This study employs three tree-based models for prediction. To enhance the generalizability of the findings, a Decision Tree, Support Vector Regression (SVR), and a Back Propagation (BP) Neural Network are further adopted for robustness testing. The results indicate that:

(a) The identification of key predictive variables remains stable. In the Decision Tree model, RD, GI, and Digital rank as the top three variables in terms of SHAP values. In the other models, the contributions of these three variables are also substantially higher than those of the remaining variables, which is consistent with the baseline XGBoost results; (b) The predictive ability of the Digital index is further validated. In the Decision Tree model, *Digital* ranks third in importance. In the SVR model, *Digital* belongs to the first tier together with RD and GI. In the BP Neural Network model, *Digital* also exhibits a positive effect. Although different models display varying sensitivities to individual variables (for example, the BP Neural Network places greater emphasis on *PSI* and *LOU*), the *Digital* index consistently appears as a positive predictive factor across all models, confirming the predictive power of digital infrastructure for regional innovation capability.

Therefore, after replacing the prediction models, the positive predictive effect of the *Digital* index remains stable, indicating that the research findings possess strong cross-model robustness.

3.3.2. Winsorization

To eliminate the potential influence of extreme values on the model prediction results, this study applies 1% and 99% winsorization to all continuous variables. Specifically, values below the 1st percentile are replaced with the 1st percentile, while values above the 99th

percentile are replaced with the 99th percentile. The model (XGBoost) is then re-estimated to examine whether the conclusions are affected by extreme observations.

The results show that:

- (a) Model predictive performance. Compared with the full-sample results (Table 4), the predictive performance of the winsorized model remains largely unchanged ($R^2 = 0.910$, RMSE = 0.362, MAE = 0.261, and MAPE = 2.480%).
- (b) Importance of predictive variables. After winsorization, the three most important variables are *GI* (0.480), *RD* (0.390), and *Digital* (0.286). Compared with the full-sample results (Table 4), the SHAP value of *Digital* changes by only 2.7% and remains ranked third in importance.

Therefore, no substantial changes are observed in either model performance or variable importance, indicating that extreme values have a negligible impact on the model estimation results and that the findings are highly robust.

3.3.3. Excluding Municipalities

To eliminate the potential influence of extreme samples from municipalities directly under the central government on the prediction results, this study re-estimates the XGBoost model after excluding the samples of Beijing, Shanghai, Tianjin, and Chongqing, thereby testing the robustness of the conclusions in Section 4.2.

The results indicate that:

- (a) Model predictive performance. Compared with the full-sample results (Table 4), the predictive performance of the model remains largely unchanged after excluding the municipalities ($R^2 = 0.934$, RMSE = 0.316, MAE = 0.223, and MAPE = 2.146%).
- (b) Importance of predictive variables. After excluding the municipalities, the three most important variables are *GI* (0.531), *RD* (0.381), and *Digital* (0.322). Compared with the full-sample results (Table 4), the SHAP value of *Digital* changes by 9.5% and remains ranked third in importance.

Therefore, no substantial changes are observed in any of the evaluation metrics, indicating that municipalities are not the key samples driving the prediction results. The findings are thus highly robust.

3.3.4. Expanding-Window Validation

To further examine the stability of the model's predictive performance, this study employs an expanding-window validation approach as a robustness check. Specifically, data from 2011–2018 are used as the initial training set and data from 2019 as the test set. The training sample is then progressively expanded by adding one additional year of data at each step, and the model is used to predict regional innovation levels in the subsequent year. In total, five expanding windows are constructed, corresponding to out-of-sample forecasts for the period 2019–2023. By comparing predictive performance across different test years, we assess whether the model maintains stable forecasting ability over time.

The results indicate that the model consistently achieves strong predictive performance across all five expanding windows, with an average R^2 value of 0.871, demonstrating robust and stable out-of-sample predictive capability across years. Moreover, the average SHAP value of the *Digital* variable is 0.307, consistently ranking among the top three most important predictors with only minor fluctuations. These findings suggest that digital infrastructure makes a persistent and stable contribution to predicting regional innovation levels. Its predictive value is not dependent on a particular sample period, thereby further confirming the robustness of the study's conclusions.

3.4. Further Analysis

This study further examines the predictive performance and importance of the *Digital* index for regional innovation capability from the perspectives of region, *Digital* level, and R&D intensity (*RD*) level. The results are reported in Table 5.

Table 5. Heterogeneity Analysis Results.

Grouping Variable	Subgroup	R ²	MAPE (%)	Digital SHAP Value	SHAP Ranking
REGION	Eastern Region	0.908	2.33%	0.172	3rd
REGION	Central and Western Regions	0.843	2.72%	0.458	2nd
<i>Digital</i>	High Group	0.841	3.44%	0.180	4th
<i>Digital</i>	Low Group	0.852	2.66%	0.454	2nd
<i>RD</i>	High Group	0.741	2.33%	0.341	1st
<i>RD</i>	Low Group	0.755	2.92%	0.384	3rd

3.4.1. Regional Heterogeneity

There are significant differences between eastern China and the central and western regions in terms of economic development, infrastructure construction, and innovation capability. To examine whether the predictive effect of *Digital* on regional innovation capability exhibits regional heterogeneity, this study follows the approach of Zhang H et al. [20] and divides the sample into two subsamples: the eastern region and the central and western regions. Model estimation and SHAP analysis are then conducted separately.

The results are presented in Table 5. In terms of SHAP-based feature importance, *Digital* exhibits a higher SHAP value and ranking in the central and western regions (SHAP = 0.458, ranked 2nd, second only to *GI*) than in the eastern region (SHAP = 0.172, ranked 3rd, behind *GI* and *RD*). This indicates that the predictive power of *Digital* for regional innovation capability is significantly stronger in the central and western regions than in the eastern region. Increasing investment in digital infrastructure in the central and western regions is likely to generate greater innovation returns and help narrow regional disparities in digital infrastructure development and innovation capability.

3.4.2. Heterogeneity by Digital Infrastructure Level

To examine whether the predictive effect of digital infrastructure varies across different development levels, the sample is divided into a high-*Digital* group and a low-*Digital* group based on the median value of *Digital*.

The results show that the importance of *Digital* is higher in the low-*Digital* group (SHAP = 0.454, ranked 2nd) than in the high-*Digital* group (SHAP = 0.180, ranked 4th). This suggests that digital infrastructure contributes the most to prediction during the early stages of digital infrastructure development. As digital infrastructure reaches a higher level of development, its marginal predictive contribution gradually becomes saturated, and the determinants of innovation capability become more diversified (e.g., *RD*, *GI*, and *PSI*). Therefore, prioritizing support for regions with relatively weak digital infrastructure may lead to greater improvements in innovation capability.

3.4.3. RD Heterogeneity

RD is an important factor in predicting regional innovation capability. To examine whether the predictive effect of *Digital* varies across different levels of R&D intensity, the sample is divided into a high-*RD* group and a low-*RD* group based on the median value of *RD*, and separate model estimations are conducted.

The results in Table 5 show that, in the high-RD group, the SHAP value of *Digital* is 0.341, ranking 1st and surpassing all other variables, making it the most important predictor of regional innovation capability. In the low-RD group, the SHAP value of *Digital* is 0.384, ranking 3rd, behind *GI* and *RD*. These findings indicate that in regions with sufficient R&D investment, digital infrastructure can generate synergistic effects with R&D activities and become a key predictor of innovation capability. In regions with relatively insufficient R&D investment, digital infrastructure can, to some extent, compensate for the lack of R&D resources.

4. Conclusions and Implications

4.1. Conclusions

This study constructs a digital infrastructure indicator system from the perspectives of digital infrastructure construction and digital infrastructure application, and employs the XGBoost model and explainable machine learning methods to predict regional innovation capability. The main findings are as follows:

- (a) The composite *Digital* index contributes to the prediction of regional innovation capability, and its predictive power is stronger than that of several commonly used external environmental characteristics, such as human capital and openness to the outside world. Among the digital infrastructure indicators, long-distance optical cable density, broadband access rate, number of domain names, mobile phone penetration rate, and Internet penetration rate are more important than other indicators in predicting regional innovation capability.
- (b) Compared with the eastern region, the Digital index exhibits stronger predictive power for regional innovation capability in the central and western regions. Compared with regions with higher levels of digital infrastructure development, the Digital index has stronger predictive power in regions with lower levels of digital infrastructure development. Compared with regions with lower R&D intensity, the *Digital* index demonstrates stronger predictive power in regions with higher R&D intensity.
- (c) Previous studies have mainly focused on the linear relationship between digital infrastructure and regional innovation capability. However, in this study, after taking the logarithm of the prediction target, both the model fit and the importance of explanatory variables are improved. Therefore, the R^2 and SHAP values of explanatory variables are affected by the numerical forms of both the explanatory and dependent variables.

Overall, the findings highlight the strategic importance of digital infrastructure as an enabling factor for innovation-driven growth, regional resilience, and sustainable development, particularly in less-developed regions.

4.2. Implications

4.2.1. Theoretical Implications

This study provides several important theoretical contributions to the existing literature.

- (a) The predictive relationship between digital infrastructure indicators and regional innovation capability exhibits significant nonlinear and stage-dependent characteristics. Based on the XGBoost and SHAP methods, this study finds that indicators such as Internet penetration rate, broadband access rate, and number of domain names contribute substantially more to the prediction of regional innovation capability once they exceed certain threshold levels. This suggests that the predictive contribution of digital infrastructure indicators varies across different value ranges and exhibits nonlinear patterns within the trained model. Therefore, traditional linear econometric models may be insufficient to capture the complex predictive relationships between

the two variables. Future research should pay greater attention to nonlinear association patterns and employ machine learning methods to identify heterogeneous predictive contributions of digital infrastructure indicators across different stages of development.

- (b) This study extends the research perspective on the relationship between digital infrastructure and regional innovation. Existing studies mainly focus on the causal effects of digital infrastructure on economic growth, green development, or innovation efficiency. From the perspective of predictive capability, this study demonstrates that digital infrastructure is an important predictor for identifying future changes in innovation capability. In particular, the Digital index consistently ranks among the top three most important predictors across multiple models. The findings are broadly consistent with patterns documented in previous studies on digital infrastructure and innovation., providing a new predictive perspective and empirical evidence for understanding the associations between digital infrastructure indicators and regional innovation capability.
- (c) This study further demonstrates the applicability and theoretical value of explainable machine learning in economic and management research. Although traditional methods such as difference-in-differences (DID) and spatial econometric models are effective in identifying causal relationships, they generally require predefined model specifications and have limited ability to capture complex nonlinear relationships and variable heterogeneity. By combining machine learning with the SHAP method, this study quantifies the magnitude and direction of predictive contributions of key variables to model outcomes, providing a clearer understanding of the predictive relationship between digital infrastructure and regional innovation capability. These findings suggest that this approach offers useful methodological insights for a wide range of empirical studies.

4.2.2. Practical Implications

Since this study focuses on predictive relationships rather than causal identification, the following implications should be interpreted as policy insights derived from predictive patterns rather than evidence of direct causal effects. This study provides several practical implications for policymakers and relevant government departments.

- (a) Policymakers should continue to monitor and support digital infrastructure development, particularly in regions with relatively weak development levels, particularly in regions with relatively weak development levels. The results indicate that indicators such as Internet penetration rate, broadband access rate, and long-distance optical cable density make substantial positive contributions to model predictors of regional innovation capability. Regions with higher levels of digital infrastructure development are generally associated with higher levels of innovation, suggesting that these indicators contain substantial predictive information regarding regional innovation capability.
- (b) Greater attention should be paid to improving the application level of digital infrastructure. This study finds that application-oriented indicators, including the number of domain names, Internet penetration rate, and mobile phone penetration rate, make larger contributions to the prediction of regional innovation capability. This suggests a strong association between the extent of digital technology adoption and regional innovation capability. From a predictive perspective, regions characterized by higher levels of digital technology adoption tend to exhibit stronger innovation-related prediction signals. Therefore, policymakers may consider not only expanding physical infrastructure but also supporting the diffusion and application of digital technologies,

promoting the diffusion of digital resources, enhancing the digital usage capabilities of residents and firms, and improving the utilization of digital resources so as to create a more favorable digital environment for regional innovation.

- (c) The coordination between digital infrastructure and R&D investment should receive greater attention. The results show that the predictive contribution of the Digital index is higher in regions with greater R&D intensity, indicating that digital infrastructure contributes more strongly to model predictions in regions with higher levels of R&D investment. Therefore, when evaluating and monitoring regional innovation potential, policymakers may consider jointly assessing digital infrastructure conditions and R&D investment levels. in order to better identify and support changes in regional innovation capability.

5. Discussion

5.1. Differences Between Machine Learning Methods and Traditional Econometric Approaches

This study employs the XGBoost model to predict regional innovation capability. The results demonstrate that the model achieves high predictive accuracy and maintains relatively stable performance after model replacement. In contrast to the findings of some previous studies [1,30], this study finds that after applying a logarithmic transformation to the dependent variable, the statistical significance of some variables decreases, while the overall predictive performance of the model improves substantially. This suggests that the results of significance testing in traditional econometric studies are not necessarily consistent with the predictive performance of machine learning models. Traditional econometric research focuses primarily on whether statistically significant relationships exist between variables, emphasizing parameter estimation and marginal effects. In contrast, machine learning prediction studies place greater emphasis on a model's fitting and generalization capabilities, with predictive accuracy and model stability serving as the primary evaluation criteria.

When the dependent variable exhibits a highly skewed distribution, logarithmic transformation can reduce the influence of extreme values on the model training process, enabling the model to better capture the overall patterns of variation among variables and thereby improving predictive stability and generalization performance. Furthermore, the SHAP results indicate that the contributions of different variables to the model's predictions vary considerably, and that several variables exhibit nonlinear association patterns within the trained model. These findings suggest that the predictive relationship captured by the machine-learning model involves complex nonlinear associations that may not be fully represented by a pre-specified linear regression specification.

By comparison, machine learning methods such as XGBoost are more capable of capturing complex nonlinear relationships and interaction effects among high-dimensional variables. When combined with the SHAP framework, these methods also allow for interpretable analysis of the predictive contributions of individual variables, thereby enhancing both model interpretability and predictive performance. This further indicates that integrating machine learning methods with explainable analysis can provide a useful predictive perspective for understanding the association between digital infrastructure and regional innovation capability.

5.2. Nonlinear Characteristics of Digital Infrastructure

Consistent with previous studies [30], the predictive relationship captured by the model between digital infrastructure and regional innovation capability exhibits pronounced stage-dependent and nonlinear characteristics. For example, the SHAP values of variables such as Internet penetration rate, broadband access rate, and long-distance

optical cable density are relatively low or even negative at lower development levels. Once these variables reach certain threshold levels, their SHAP values increase rapidly. However, at higher values, the growth of their SHAP values gradually stabilizes.

One possible interpretation of this pattern is that the model captures different association structures across levels of digital infrastructure development. At lower levels of digital infrastructure indicators, the model assigns relatively small or negative SHAP values. Consequently, the model captures relatively weak predictive contributions in this range. As digital infrastructure indicators increase, the SHAP values become more positive, suggesting stronger predictive contributions within the trained model. However, at relatively high values, the upward trend of SHAP values gradually flattens. This pattern may indicate that other predictors also contribute substantially to model predictions at higher levels of digital infrastructure, resulting in a stabilization of SHAP values.

These findings further suggest that traditional linear models may have difficulty capturing nonlinear changes in predictive contributions across different ranges of explanatory variables. In contrast, machine learning models such as XGBoost can automatically capture nonlinear relationships and interaction effects among variables through their tree-based structures, thereby providing a more accurate characterization of the predictive role of digital infrastructure.

5.3. Heterogeneity Characteristics

The heterogeneity analysis indicates that the predictive importance of digital infrastructure for regional innovation capability varies significantly across regions and development stages. Specifically, the Digital index exhibits greater importance in the central and western regions and in regions with relatively low levels of digital infrastructure development. Moreover, its predictive contribution is more pronounced in regions with higher levels of R&D investment. These findings suggest that the predictive relationship captured by the model varies across regional development conditions and stages of digitalization, which is consistent with previous studies [7,18,19,31].

From the perspective of regional differences, digital infrastructure in eastern China is generally more mature, with indicators such as Internet penetration and broadband access remaining at relatively high levels for a long period. As a result, these indicators exhibit relatively limited variation, which may contribute to the lower importance of the Digital index in the prediction model. In contrast, digital infrastructure in the central and western regions is still undergoing rapid development, and the corresponding indicators exhibit greater variation. Consequently, digital infrastructure indicators contribute more strongly to model predictions in these regions. Similarly, in regions with relatively low levels of digital infrastructure development, the importance of digital infrastructure variables is also higher, indicating that variation in digital infrastructure indicators is more readily reflected in the model's predictions during the early stages of digitalization, when their predictive contribution to model outcomes is more evident.

In addition, the importance of the Digital index increases further in regions with high levels of R&D investment, suggesting that digital infrastructure indicators contribute more strongly to model predictions in areas with more intensive R&D activities. This finding indicates that the predictive importance of digital infrastructure varies with regional R&D conditions and stages of development. Compared with traditional aggregate regression analysis, machine learning models are better able to identify differences across subsamples, thereby enhancing the ability to capture heterogeneous characteristics of regional innovation capability.

5.4. Limitations

This study still has several limitations.

- (a) This study is a machine learning-based prediction study that focuses on the predictive power of digital infrastructure for regional innovation capability. Therefore, the findings reflect the predictive associations between explanatory variables and the target variable and cannot directly explain the specific mechanisms through which digital infrastructure influences regional innovation capability.
- (b) This study employs provincial-level panel data, which are suitable for capturing macro-level regional differences. However, such data make it difficult to further characterize micro-level features at the city or firm level.
- (c) Although the XGBoost and SHAP methods are effective in identifying nonlinear relationships and quantifying the predictive contributions of variables, they have certain limitations in explaining underlying mechanisms. Future research could further combine spatial econometric methods, causal inference approaches, and micro-level data analysis to provide a deeper understanding of the relationship between digital infrastructure and regional innovation capability.
- (d) The dependent variable considered in this study is measured solely by the number of regional patent grants, which may not fully capture the quality and efficiency of innovation. Future research could incorporate more comprehensive indicators, such as high-value patents and innovation output efficiency, to better evaluate the predictive role of digital infrastructure in regional innovation. In terms of digital infrastructure measurement, due to data availability constraints, this study focuses only on network construction and digital technology application. Future studies may further incorporate indicators related to emerging digital infrastructures, such as data centers, cloud computing, and artificial intelligence, to provide a more comprehensive assessment.

Author Contributions: Conceptualization, S.X. (Shasha Xie) and S.X. (Shusheng Xu); Methodology, S.X. (Shasha Xie) and S.X. (Shusheng Xu); Software, W.X.; Validation, W.X. and P.Z.; Formal analysis, W.X.; Investigation, X.C. and P.Z.; Resources, J.W., M.X. and X.C.; Data curation, S.X. (Shusheng Xu), M.X. and P.Z.; Writing—original draft, S.X. (Shusheng Xu); Writing—review & editing, S.X. (Shusheng Xu); Visualization, Y.L.; Supervision, G.Y., Y.L. and J.W.; Project administration, G.Y., Y.L. and P.Z.; Funding acquisition, G.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Project of Hubei Key Laboratory of Basin Water Security (grant number QTKS0034W2581); the Science Research Fund Project of Wuhan Institute of Technology (grant numbers K2024023, K2024025); General Project of Hubei Provincial Social Science Fund (grant number HBSKJJ20253258); the Graduate Innovative Fund of Wuhan Institute of Technology (grant number CX2025366); and Philosophy and Social Science Research Project of colleges and Universities in Hubei Province (Grant number 24Y053) The APC was funded by the corresponding research projects above.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The raw statistical indicators are derived from publicly published official statistical yearbooks, which can be freely accessed by all researchers via official channels. The compiled panel dataset sorted for this paper contains integrated processed indicators, and we provide the sorted dataset upon reasonable request to avoid repeated data collation workload for other scholars.

Conflicts of Interest: Author Wei Xu was employed by the company Meizhou Geological Survey Center of Guangdong Geological Bureau, Meizhou, China. Author Jianqiu Wu was employed by the company China Construction Eighth Engineering Bureau Co., Ltd. Author Minghua Xiao was employed by the company China Railway 11th Engineering Bureau Group Co., Ltd. Author Peng Zhang was employed by the company China Railway Design Group Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Appendix A

Table A1. Entropy Weight Method Weights of Eight Variables.

Variables	<i>LDOCD</i>	<i>BAR</i>	<i>MPEC</i>	<i>DOM</i>	<i>WEB</i>	<i>MPPR</i>	<i>IPR</i>	<i>PC</i>
Weights	0.0685	0.0533	0.0478	0.2379	0.4596	0.03375	0.0546	0.04433

Table A2. Results of Pearson Correlation Coefficient Matrix and Variance Inflation Factor (VIF) for Eight Variables.

Variables	<i>LDOCD</i>	<i>BAR</i>	<i>MPEC</i>	<i>DOM</i>	<i>WEB</i>	<i>MPPR</i>	<i>IPR</i>	<i>PC</i>
<i>LDOCD</i>	1.0000	0.2845	0.1490	0.3005	0.2458	0.4426	0.2359	0.3983
<i>BAR</i>	0.2845	1.0000	0.5903	0.3698	0.3056	0.7819	0.9204	0.6860
<i>MPEC</i>	0.1490	0.5903	1.0000	0.5225	0.5935	0.7039	0.4510	0.6644
<i>DOM</i>	0.3005	0.3698	0.5225	1.0000	0.8525	0.5987	0.2295	0.5881
<i>WEB</i>	0.2458	0.3056	0.5935	0.8525	1.0000	0.5967	0.1452	0.6644
<i>MPPR</i>	0.4426	0.7819	0.7039	0.5987	0.5967	1.0000	0.6675	0.8386
<i>IPR</i>	0.2359	0.9204	0.4510	0.2295	0.1452	0.6675	1.0000	0.5359
<i>PC</i>	0.3983	0.6860	0.6644	0.5881	0.6644	0.8386	0.5359	1.0000
VIF	1.3811	11.208	2.4906	4.1102	5.5123	5.9756	7.8722	4.4917

Table A3. Summary of Model Hyperparameters, Tuning Procedures, Software, and Validation Strategies.

Item	XGBoost	Random Forest	LightGBM
Core Fixed Hyperparameters	n_estimators = 300 max_depth = 6 learning_rate = 0.05 subsample = 0.7 colsample_bytree = 0.7 reg_alpha = 0.5 reg_lambda = 1.0 eval_metric = rmse random_state = 42 verbosity = 0	n_estimators = 200 max_depth = 10 min_samples_split = 5 min_samples_leaf = 2 random_state = 42 n_jobs = −1	n_estimators = 200 max_depth = 6 learning_rate = 0.05 subsample = 0.7 colsample_bytree = 0.7 reg_alpha = 0.5 reg_lambda = 1.0 random_state = 42 verbose = −1
Parameter Tuning Procedure	No grid search or random search was performed; fixed parameters were adopted. Early stopping was enabled during training with a patience of 20 rounds, and the model with the optimal iteration was saved	No parameter tuning was performed; all parameters were set to fixed values	No parameter tuning was performed; all parameters were set to fixed values

Table A3. Cont.

Item	XGBoost	Random Forest	LightGBM
Standardized Feature Usage	No	No	No
Validation Strategy/Dataset Split	Time-based stratified split: Training set: year $\leq$ 2021 Validation set: year = 2022 (exclusively for early stopping) Test set: year = 2023 No k-fold cross-validation was applied; time-series extrapolation prediction was performed	Time-based stratified split: Training set: year $\leq$ 2021 Test set: year = 2023 Samples from 2022 were only used as the validation set for XGBoost early stopping	Same as the time-based split above
Training Early Stopping Configuration	Early Stopping enabled; training stops if no improvement is observed for 20 rounds, and the optimal iteration model is automatically saved	Early stopping not enabled	Early stopping not enabled
Dependent Software & Packages	Python, pandas/numpy/matplotlib, xgboost, scikit-learn, shap	Python, pandas/numpy/matplotlib, scikit-learn, shap	Python + pandas/numpy/matplotlib/lightgbm, scikit-learn, shap

References

- Ma, R.; Lin, B. Digital infrastructure construction drives green economic transformation: Evidence from Chinese cities. *Humanit. Soc. Sci. Commun.* **2023**, *10*, 460. [\[CrossRef\]](#)
- Nie, J.; Shen, J.; Ren, X. Digital Infrastructure, New Digital Infrastructure, and Urban Carbon Emissions: Evidence from China. *Atmosphere* **2025**, *16*, 199. [\[CrossRef\]](#)
- Zhang, F.; Wang, Y.; Liu, W. Science and Technology Resource Allocation, Spatial Association, and Regional Innovation. *Sustainability* **2020**, *12*, 694. [\[CrossRef\]](#)
- Dan, C.; Jia, Z. How does digital infrastructure promote regional innovation. *CSS* **2025**, *40*, 118–127. [\[CrossRef\]](#)
- Nie, C.; Zhong, Z.; Feng, Y. Can digital infrastructure induce urban green innovation? New insights from China. *Clean Technol. Environ. Policy* **2023**, *25*, 3419–3436. [\[CrossRef\]](#)
- Yang, Y.; Chen, W.; Gu, R. How does digital infrastructure affect industrial eco-efficiency? Considering the threshold effect of regional collaborative innovation. *J. Clean. Prod.* **2023**, *427*, 139248. [\[CrossRef\]](#)
- Du, Z.Y.; Wang, Q. Digital infrastructure and innovation: Digital divide or digital dividend? *J. Innov. Knowl.* **2024**, *9*, 100542. [\[CrossRef\]](#)
- Zhai, C.; Ding, X.; Zhang, X.; Jiang, S.; Zhang, Y.; Li, C. Assessing the Effects of Urban Digital Infrastructure on Corporate Environmental, Social and Governance (ESG) Performance: Evidence from the Broadband China Policy. *Systems* **2023**, *11*, 515. [\[CrossRef\]](#)
- Chao, X.; Lian, Y.; Luo, L. Impact of new digital infrastructure on high-quality development of manufacturing. *Financ. Trade Res.* **2021**, *32*, 1–13.
- Meng, G.; Wang, K.; Wang, F.; He, H. Spatial-temporal Pattern and Mechanism of the Coupling Coordination Relationship Between Digital Economy and Urban-rural Integration Development at Prefectural level in the Yangtze River Economic Belt. *Econ. Geogr.* **2024**, *44*, 11.
- He, X.; Jia, D.; Shi, C. The coupled and coordinated relationship between digital infrastructure construction and tourism economic resilience in the Yangtze River Economic Belt. *J. Shaanxi Norm. Univ. (Nat. Sci. Ed.)* **2026**, *54*, 107–121.
- Wang, Q.; Li, G.; Du, M.; Zhou, X.; Liang, J. The impact of new digital infrastructure on green total factor productivity. *Front. Energy Res.* **2024**, *12*. [\[CrossRef\]](#)
- Liang, K.; Liu, W. Digital infrastructure and urban-rural income gap: Empirical evidence from China. *Int. Rev. Econ. Financ.* **2026**, *107*, 105003. [\[CrossRef\]](#)

14. Che, S.; Wen, L.; Wang, J. Global insights on the impact of digital infrastructure on carbon emissions: A multidimensional analysis. *J. Environ. Manag.* **2024**, *368*, 122144. [\[CrossRef\]](#)
15. Kafourous, M.I. The impact of the Internet on R&D efficiency: Theory and evidence. *Technovation* **2006**, *26*, 827–835. [\[CrossRef\]](#)
16. Zhao, X.; Liu, J. The impact of digital infrastructure on labor productivity. *Econ. Anal. Policy* **2025**, *86*, 274–287. [\[CrossRef\]](#)
17. Xu, X.; Watts, A.; Reed, M. Does access to internet promote innovation? A look at the U.S. broadband industry. *Growth Change* **2019**, *50*, 1423–1440. [\[CrossRef\]](#)
18. Guo, B.; Hu, P.; Lin, J. The effect of digital infrastructure development on enterprise green transformation. *Int. Rev. Financ. Anal.* **2024**, *92*, 103085. [\[CrossRef\]](#)
19. Tian, X.; Lu, H. Digital infrastructure and cross-regional collaborative innovation in enterprises. *Financ. Res. Lett.* **2023**, *58*, 104635. [\[CrossRef\]](#)
20. Zhang, H.; Zhang, K.; Yan, T.; Cao, X. The impact of digital infrastructure on regional green innovation efficiency through industrial agglomeration and diversification. *Humanit. Soc. Sci. Commun.* **2025**, *12*, 220. [\[CrossRef\]](#)
21. Li, C.; Wen, M.; Jiang, S.; Wang, H. Assessing the effect of urban digital infrastructure on green innovation: Mechanism identification and spatial-temporal characteristics. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 320. [\[CrossRef\]](#)
22. Zhang, H.; Wang, Y.; Chen, Y.; Wang, W. How does digital infrastructure influence synergistic effect on reducing pollution and carbon emissions? A new method based on double machine learning. *J. Clean. Prod.* **2025**, *516*, 145848. [\[CrossRef\]](#)
23. Fang, H. A causal inference for digital infrastructure construction and green innovation using double machine learning. *Discov. Comput.* **2026**, *29*, 32. [\[CrossRef\]](#)
24. Huang, J.; Xiang, Y.; Duan, X.; Du, M. Digital Infrastructure Construction and Energy Efficiency: A Quasi-Natural Experiment with Double Machine Learning. *Emerg. Mark. Financ. Trade* **2026**, *62*, 2646–2660. [\[CrossRef\]](#)
25. Ghoddusi, H.; Creamer, G.G.; Rafizadeh, N. Machine learning in energy economics and finance: A review. *Energy Econ.* **2019**, *81*, 709–727. [\[CrossRef\]](#)
26. Pyo, S.; Choi, S.O. Regional innovation and economic growth: Empirical insights from FGLS, FE-DKSE, and XGBoost-SHAP approach. *J. Open Innov. Technol. Mark. Complex.* **2025**, *11*, 100524. [\[CrossRef\]](#)
27. Ji, T.; Li, K.; Sun, Q.; Duan, Z. Urban transport emission prediction analysis through machine learning and deep learning techniques. *Transp. Res. Part D Transp. Environ.* **2024**, *135*, 104389. [\[CrossRef\]](#)
28. Xie, H.; Xu, X.; Yan, F.; Qian, X.; Yang, Y. Deep Learning for Multi-Country GDP Prediction: A Study of Model Performance and Data Impact [Internet]. *arXiv* **2024**. [\[CrossRef\]](#)
29. Lundberg, S.; Lee, S.I. A Unified Approach to Interpreting Model Predictions [Internet]. *arXiv* **2017**. [\[CrossRef\]](#)
30. Sun, H.; Gu, T.; Ma, S. Digital infrastructure and regional innovation disparities in China: A core-periphery analysis with policy implications. *Appl. Econ.* **2026**, 1–18. [\[CrossRef\]](#)
31. Hu, R.; Bao, Z.; Lin, Z.; Lv, K. The Innovative Construction of Provinces, Regional Artificial Intelligence Development, and the Resilience of Regional Innovation Ecosystems: Quasi-Natural Experiments Based on Spatial Difference-in-Differences Models and Double Machine Learning. *Sustainability* **2024**, *16*, 8251. [\[CrossRef\]](#)

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