Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 7, No. 1, 2026

A Multi-Stage Modeling Framework Integrating Random Forests and Cluster Analysis: Feature Identification and Prediction for the Global Innovation Index

Yifei Guo *

Central University of Finance and Economics, Beijing, China

* **Corresponding author**: (Email: 1930726348@qq.com)

Abstract: Based on data from the Global Innovation Index, this paper constructs an integrated data-driven framework that combines feature selection, cluster analysis, and predictive modeling. First, a random forest model is used to assess the importance of multidimensional indicators, identifying key variables from high-dimensional features to effectively capture complex nonlinear relationships. Second, the K-means method is employed to stratify the innovation levels of different countries, achieving structured classification. Building on this foundation, a sliding-window-based random forest regression model is introduced to transform short-sequence data into supervised learning samples, enabling high-precision forecasting of future indicators. The model demonstrates strong fitting capability and stability in testing. The core innovation of this study lies in the systematic integration of feature selection, structural classification, and predictive modeling into a unified analytical workflow, thereby enhancing the model's adaptability to complex data and its interpretability. This method avoids reliance on a single model and demonstrates good generalization performance under various data structural conditions, making it applicable to scenarios involving multi-indicator comprehensive evaluation and dynamic forecasting. Overall, the proposed model framework enhances forecasting accuracy while simultaneously providing a synergistic characterization of key drivers and structural features, offering a universal and efficient technical approach for the analysis of complex systems.

Keywords: Random Forest, K-means Clustering, Global Innovation Index.

1. Introduction

Against the backdrop of intensifying global competition and rapid technological advancements, characterizing the features of complex systems and achieving effective predictions has become a critical research issue. Multi-indicator data typically exhibits characteristics such as high dimensionality, nonlinearity, and significant structural variations, making it difficult for traditional single-method approaches to balance interpretability and predictive accuracy. This issue is particularly pronounced in cross-entity comparisons and dynamic analyses.

Existing methods largely rely on linear models or single algorithms, often suffering from limitations such as insufficient feature selection, limited utilization of structural information, and weak predictive capabilities for short time series. On the one hand, redundant information in high-dimensional data can compromise model stability; on the other hand, the lack of characterization of sample hierarchical structures makes it difficult for analysis results to reflect overall differences. Furthermore, traditional time series methods perform poorly when data length is limited, making it difficult to meet practical needs [1].

To address these issues, this paper proposes a multi-stage collaborative algorithmic framework. First, a random forest is employed to screen for key features, enhancing the effectiveness of variable selection; second, the K-means method is integrated to perform structural partitioning, characterizing features across different categories; building on this foundation, a sliding-window-based ensemble model is introduced to achieve stable predictions of future trends. This method integrates the strengths of multiple models within a unified framework, avoiding the pitfalls of fragmented application [2, 3].

Using Global Innovation Index data as an example, this framework can simultaneously identify key factors, perform structural segmentation, and forecast trends, demonstrating good applicability and scalability, and providing an efficient modeling approach for complex data analysis.

2. Data Processing and Integrated Modeling Framework

2.1. Data Source and Preprocessing

The analysis is based on the Global Innovation Index (GII) dataset released by the World Intellectual Property Organization (WIPO), covering the period from 2022 to 2025. The dataset includes 536 annual observations from 141 countries and regions, with each record containing the GII overall score along with a large set of sub-indicators [4, 5].

Data preprocessing is carried out to improve data quality and ensure comparability across countries. Indicators with more than 30% missing values are removed. For the remaining variables, missing values are filled using a combination of country-level mean values and global median values, so that both national characteristics and overall distribution patterns are preserved.

To reduce the influence of extreme observations, all variables are Winsorized at the 1% and 99% levels. In addition, Z-score normalization is applied at the country level to eliminate scale differences across indicators and facilitate subsequent analysis.

After preprocessing, 100 valid features are retained for further analysis.

2.2. Methodology

The study adopts a data-driven approach that combines feature selection, clustering analysis, and forecasting methods to analyze the GII dataset.

(1) Random Forest for Feature Selection

A random forest regression model is used to identify the most important factors affecting national innovation performance. The model can handle nonlinear relationships and provides a ranking of feature importance.

Feature importance is calculated based on permutation importance using out-of-bag (OOB) samples:

$$Importance(X_j) = \frac{1}{B}\sum_{b=1}^{B}\left(err_{OOB_b}^{perm} - err_{OOB_b}\right) \quad (1)$$

Where OOB_b represents the out-of-bag samples of the b-th tree, err_{OOB_b} is the prediction error on the original data, and $err_{OOB_b}^{perm}$ is the error after randomly permuting feature X_j.

The model is trained using cross-sectional data from 2025 with 500 trees. Based on the importance scores, the top 20 features are selected as the core variables for further analysis.

(2) K-means Clustering

K-means clustering is applied to group countries into different innovation levels. The objective is to minimize the within-cluster sum of squares:

$$J(C) = \sum_{i=1}^{k}\sum_{x\in C_i}|x-\mu_i|^2 \quad (2)$$

Where μ_i is the centroid of cluster C_i.

Clustering is performed based on the GII overall scores in 2025, which provides a direct classification of countries according to their innovation performance. The number of clusters is set to $k=3$, corresponding to high, medium, and low innovation groups. Multiple initializations are used to improve the stability of the results.

(3) Time Series Forecasting

To predict the GII scores for 2026, a random forest-based regression model is used.

Due to the limited time span of the data, a sliding window approach is adopted. Historical observations are transformed into feature vectors by including lagged values and growth rates. This allows short time series to be converted into multiple supervised learning samples, resulting in 128 observations.

The dataset is split into training and testing sets with a ratio of 8:2. Model performance is evaluated using R^2, MAPE, and MSE. The model achieves strong performance on the test set, with $R^2 = 0.9897$ and $MAPE = 4.04\%$.

Finally, the trained model is applied to the most recent data (2023–2025) to generate predictions for 2026.

3. Model Results and Algorithmic Analysis

3.1. Identification of Key Innovation Drivers

Using cross-sectional data from 2025, the random forest model identifies the most influential determinants of national innovation performance. Among the 100 candidate variables, the top 20 features are selected based on importance scores. The five most significant features are summarized in Table 1.

Table 1. Comparison of power load forecasting of 403 line

Rank	Feature	Importance Score	Associated Dimension
1	Youth Population Ratio	0.5917	Human Capital and Research
2	Product of Top Three QS University Rankings	0.0897	Higher Education Quality and Talent Development
3	Number of Knowledge Workers	0.0415	Business Sophistication – Knowledge Workers
4	Patent Families (PPP-adjusted)	0.0277	Knowledge and Technology Outputs – Knowledge Creation
5	Mobile App Development (PPP-adjusted)	0.0211	Creative Outputs – Online Creativity

The results indicate that the proportion of youth population exhibits a dominant effect, with an importance score substantially higher than all other variables. This suggests that demographic structure plays a critical role in shaping innovation capacity. A higher share of young population (typically aged 15–34) implies a larger pool of potential researchers, entrepreneurs, and skilled labor with strong learning capacity and innovation propensity. This demographic advantage contributes to the expansion of R&D personnel and STEM graduates, thereby enhancing knowledge creation (e.g., patents and publications) and creative outputs (e.g., digital applications), which ultimately improves the GII score.

The product of the top three QS university rankings serves as a proxy for the quality of higher education and research capacity. This indicator reflects the strength of leading academic institutions in terms of scientific output, reputation, and international collaboration. Countries with highly ranked universities tend to have stronger research-industry linkages and are more capable of attracting global talent, which further supports innovation performance.

The number of knowledge workers captures the scale of the innovation-oriented labor force, including professionals engaged in R&D, technical services, and creative industries. In addition, patent family counts adjusted by purchasing power parity (PPP) measure the international scope of intellectual property protection relative to economic size, reflecting the efficiency of knowledge production and protection systems. Similarly, mobile application development per unit of economic scale represents digital creative output capacity, indicating the activity level of software and internet-based innovation sectors.

Figure 1 further illustrates the importance ranking of the top 15 features. In addition to the top five variables, other

influential factors include tertiary education enrollment, knowledge impact, PCT patent applications by origin, intellectual property payments, and enterprise R&D expenditure. These results show that both human capital and knowledge transformation capacity jointly shape national innovation performance.

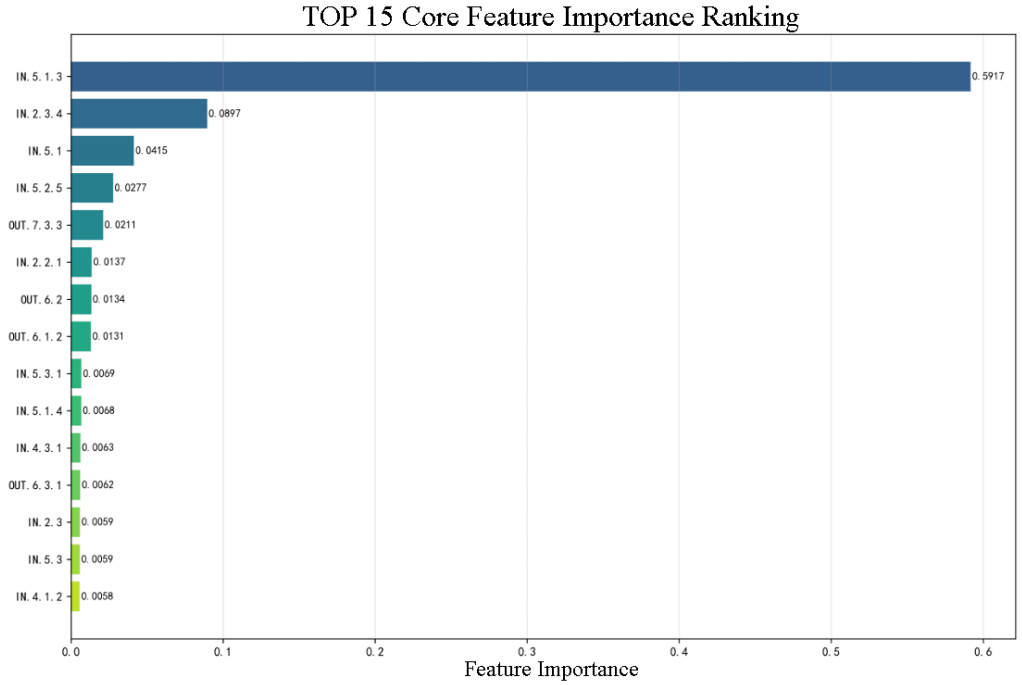

Figure 1. Ranking of the top 15 core features by importance in the GII score

These results reflect that national innovation performance is jointly shaped by demographic structure, knowledge infrastructure, and the efficiency of knowledge transformation.

3.2. Clustering Results Based on GII Scores

Based on the GII overall scores in 2025, countries are grouped into three distinct clusters representing high, medium, and low innovation levels. The clustering results are summarized in Table 2.

Table 2. Characteristics of country clusters based on the 2025 GII total score

Innovation Level	Number of Countries	Mean GII Score	Std. Dev.	GII Range (Min–Max)	Representative Countries
High Innovation Level	31	52.26	6.19	43.86 – 65.96	Switzerland, Sweden, United States, Singapore
Medium Innovation Level	50	32.50	4.60	26.34 – 42.04	Italy, Spain, Czech Republic, Malaysia
Low Innovation Level	58	19.52	3.70	11.95 – 25.88	Most African countries and some developing Asian countries

The results reveal a clear hierarchical structure of global innovation. The high-innovation group, consisting of 31 countries (22.3%), is dominated by developed economies in Europe, North America, and East Asia. This group exhibits a high average GII score (52.26) and relatively large internal variation, indicating both strong performance and diversity within advanced innovation systems.

The medium-innovation group includes 50 countries, mainly emerging and developing economies, with moderate performance levels. In contrast, the low-innovation group, comprising 58 countries (41.7%), shows significantly lower average scores (19.52) and relatively concentrated distributions. This reflects that these countries face substantial challenges in catching up in terms of innovation capacity.

Overall, the clustering results suggest a "pyramidal" structure in the global innovation landscape, where a small number of leading economies occupy the top tier, while a large proportion of countries remain at lower levels of innovation development.

3.3. Temporal Dynamics and Forecasting

This section combines a quantitative analysis of historical trends with model-based forecasts to examine the evolution of global innovation and its short-term trajectory.

(1) Historical Trends

Over the period 2022–2025, global innovation growth shows a pattern of overall slowdown accompanied by increasing divergence across countries. The average compound annual growth rate (CAGR) of GII scores is approximately 0.68%, indicating a relatively stagnant global trend.

This slowdown is consistent with declining growth in global R&D investment and increasing uncertainty in the economic environment. Growth remains uneven across sectors, with only a few high-tech industries maintaining momentum. In addition, global venture capital activity shows signs of concentration rather than broad expansion.

Despite the overall slowdown, a group of latecomer economies demonstrates relatively strong growth. Countries such as Guinea, Burundi, and Nigeria achieve growth rates

exceeding 7%, indicating that innovation catch-up remains feasible under specific conditions. At the same time, several middle-income economies continue to improve their positions in the global ranking, reflecting a gradual shift in the innovation landscape [6].

Another notable trend is the increasing spatial concentration of innovation activities. Innovation output is becoming more concentrated in a limited number of global hubs, supported by dense networks of universities, research institutions, and technology firms. This concentration effect reinforces disparities between leading and lagging regions.

Table 3 summarizes representative growth patterns across selected countries, illustrating different innovation trajectories, including steady leadership, rapid catch-up, and structurally driven growth.

Table 3. Growth Trajectories and Driving Factors of Representative Countries (2022–2025)

Country /Region	GII Score (2025)	CAGR (2022–2025)	Growth Characteristics and Key Drivers
China	56.56	High	Systematic upgrading. Innovation input (ranked 19th) and output (ranked 5th) improve simultaneously, forming an efficient innovation cycle. Key drivers include large-scale R&D investment and workforce, efficiency gains from digital technologies such as AI, the agglomeration effect of 24 top global innovation clusters, and industrial advantages in AI, semiconductors, and green technologies.
Nigeria	21.08	Very High	Breakthrough-driven catch-up. A typical latecomer pattern characterized by rapid progress driven by strong performance in unicorn valuations and venture capital attraction, reflecting high entrepreneurial activity in the digital economy.
Switzerland	65.96	Low and Stable	Stable leadership. Long-term top ranking with growth entering a mature plateau stage. Its strength lies in a highly developed and balanced innovation ecosystem with no significant weaknesses across key dimensions.
India	38.20	Steady Growth	Service- and digital-driven growth. Key drivers include ICT service exports, active venture capital markets, and strong startup financing (particularly in later-stage investments), enabling sustained leadership among lower-middle-income economies.

(2) Forecasting Results for 2026

A random forest-based forecasting model is used to generate GII predictions for 2026 across 128 countries. The model demonstrates strong predictive performance, with $R^2 = 0.9897$ and $MAPE = 4.04\%$, indicating reliable approximation of historical patterns.

The forecasting results reveal three main trends.

First, global innovation is expected to exhibit moderate acceleration. The average projected growth rate is approximately 1.86%, slightly higher than the historical average but still within a relatively stable range. This suggests a potential recovery driven by the diffusion of general-purpose technologies, although rapid expansion is unlikely.

Second, competition among leading economies is intensifying. The gap between top-ranked countries is narrowing, indicating a more competitive landscape. Switzerland is projected to remain at the top, although with a slight decline in score. Meanwhile, the United States, China, and Germany maintain stable or moderately increasing performance, as shown in Table 4.

Table 4. Forecasted GII scores and rankings of selected countries in 2026

Country	GII Score (2025)	Predicted GII Score (2026)	Growth Rate (2025–2026, %)	Predicted Rank (2026)
Switzerland	65.96	59.37	-10.00	1
United States	61.69	58.75	-4.76	2
Sweden	62.58	58.75	-6.11	3
Germany	55.46	57.36	3.43	10
China	56.56	57.17	1.09	11

The projected adjustments among top economies may reflect shifts in technological diffusion, industrial structure, and investment cycles. For example, advanced manufacturing and digital transformation contribute to Germany's growth, while China's performance continues to benefit from large-scale innovation ecosystems and integration of emerging technologies.

Third, latecomer economies are expected to sustain relatively high growth rates. Countries that have recently demonstrated strong performance are likely to continue improving, indicating that innovation-driven development remains a viable pathway for economic advancement.

(3) Discussion

The results show that the primary source of differentiation in global innovation has shifted from input scale to system efficiency. Countries with relatively moderate input levels can achieve high output performance if innovation systems effectively support knowledge transformation and commercialization.

At the same time, a structural gap remains between innovation potential and realized performance. While many countries exhibit similar foundational conditions, their ability to translate these into measurable outputs varies significantly. This highlights the importance of institutional coordination and efficient resource allocation.

It is also necessary to recognize the limitations of model-based forecasting. The predictions rely on the assumption that historical patterns will persist, while real-world outcomes may be affected by technological disruptions, geopolitical events, or macroeconomic shocks. Therefore, the forecasting results should be interpreted as scenario-based projections rather than deterministic outcomes.

This transition highlights a fundamental shift in global innovation from scale-driven expansion to efficiency-oriented development.

4. Conclusions and Methodological Implications

This study provides a data-driven analysis of the Global Innovation Index (GII) and reveals a structural shift in global innovation competition—from a focus on input scale to the

efficiency and resilience of innovation systems. Three main findings emerge.

4.1. Conclusions

First, the ability to absorb and transform knowledge is a key determinant of innovation performance. The identified core features suggest that innovation outcomes depend not only on knowledge creation but also on the effectiveness of translating research outputs into economic value. Regions with strong linkages between research, capital, and industry tend to achieve higher levels of innovation output.

Second, a gap exists between innovation potential and realized performance. Although many countries share similar structural conditions in terms of human capital and institutional frameworks, their actual innovation outcomes differ significantly. This indicates that institutional coordination and the efficiency of resource allocation play a critical role in bridging this gap.

Third, global innovation growth shows a pattern of overall slowdown combined with structural opportunities. While leading economies experience intensified competition and narrowing gaps, some latecomer countries achieve relatively high growth by focusing on specific sectors such as digital technologies and green industries. This shows that targeted strategies can enable partial catch-up in the global innovation system.

4.2. Policy Implications

Based on these findings, differentiated policy strategies are necessary for countries at different stages of innovation development.

For high-innovation countries, the focus should shift from frontier expansion to enhancing system resilience. This includes strengthening collaboration among universities, research institutions, and industries, promoting efficient mechanisms for technology transfer, and supporting long-term investment in emerging technologies. In addition, talent development systems and innovation service platforms should be improved to facilitate resource allocation and sustain innovation capacity.

For middle-innovation countries, the priority lies in improving the conversion efficiency of existing innovation resources. Policy efforts should focus on removing institutional barriers to technology commercialization, identifying key bottlenecks in the innovation system, and developing specialized industrial clusters based on existing comparative advantages.

For low-innovation countries, the emphasis should be placed on building fundamental capabilities and integrating into global innovation networks. This includes investing in basic education, digital infrastructure, and vocational training, as well as attracting external innovation resources through international collaboration and technology transfer.

In summary, global innovation competition is increasingly shaped by the efficiency of innovation ecosystems rather than the scale of inputs. Countries need to align their policy strategies with their structural conditions and focus on improving the effectiveness of knowledge transformation to achieve sustainable competitiveness.

5. Conclusions

This paper proposes a multi-stage algorithmic framework that integrates feature selection, structural partitioning, and dynamic forecasting to characterize the intrinsic patterns of complex indicator systems and achieve high-precision predictions. First, based on a feature importance evaluation mechanism using random forests, core features are selected from the original 100 variables, significantly reducing data redundancy and enhancing the model's interpretability; Second, K-means clustering is employed to partition the study subjects into different hierarchical levels, effectively revealing the distribution characteristics and differences within the overall structure; then, an ensemble regression model constructed using a sliding window achieves stable predictions under limited sample conditions, with test results showing high fitting accuracy (R^2 reaching 0.9897 and MAPE at 4.04%); Finally, this framework enables the collaborative operation of multiple models, enhancing the method's versatility and scalability while maintaining accuracy.

Overall, this method combines interpretability with predictive capabilities in the analysis of complex systems, providing an effective tool for multi-indicator evaluation and trend analysis. Future research could further explore the model's applicability in larger-scale data and multi-source heterogeneous data environments.

References

[1] Cui Youxiang, Hu Xinghua, Liao Juan, et al. Research on the Measurement and Evaluation System for Implementing the Innovation-Driven Development Strategy [J]. Science and Technology Management, 2013, 34(S1): 308–314+338. DOI: 10.19571/j.cnki.1000-2995.2013.s1.045.

[2] Su Jin. User Segmentation Based on the K-means Clustering Algorithm [J]. Digital Communications World, 2021, (06): 127-128+124.

[3] Shi Ce. A Study on User-Oriented Value in Internet Enterprises [D]. Shanghai Normal University, 2025. DOI:10.27312/d.cnki.gshsu. 2025.001393.

[4] Jia Ru. A Discussion on the Applicability of the Global Innovation Index to China: An Analysis Based on Measurement Content, Data, and Methods [J]. Research on Science, Technology Innovation and Development Strategy, 2025, 9(05): 54-66.

[5] Qi Su, Liu Lichun. Analysis of the Current Status and Influencing Factors of China's Innovation Capabilities Based on the Global Innovation Index [J]. Science and Technology Progress and Policies, 2018, 35(18): 1-10.

[6] Wang Zongjun, Wang Xue, Jiang Zhenyu. A Study on the Hierarchical Structure Model and Improvement Pathways of China's Global Innovation Index [J]. Complex Science Management, 2020, (02): 35-50.