World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3091

MEASURING DIGITAL GREEN INNOVATION IN ENTERPRISES: A LIGHTWEIGHT TEXT MINING APPROACH BASED ON ANNUAL REPORTS

QiWen Wei, ZiJie Xu*
Hainan International College, Minzu University of China, Lingshui 572400, Hainan, China.
**Corresponding Author: ZiJie Xu*

Abstract: Digital green innovation has become a critical pathway for Chinese enterprises to achieve sustainable development amid the convergence of the digital economy and green transition strategies. However, existing studies predominantly rely on structured indicators such as patents and financial data, which may fail to capture the strategic and contextual dimensions of firms' innovation activities reflected in corporate disclosures. To address this gap, this study analyzes the annual reports of 1,285 Chinese listed manufacturing firms using web scraping and text mining techniques. Based on recent literature, a multidimensional indicator system containing 11 digital green innovation keywords is constructed to measure firms' innovation activities. Empirical results reveal significant heterogeneity in keyword distribution across firms, with "artificial intelligence" and "Internet" appearing most frequently. Further analysis shows that heavily polluting enterprises exhibit significantly higher levels of green innovation, suggesting strong policy and regulatory incentives. By providing a scalable, text-based measurement framework, this study not only enriches the methodological toolkit for assessing digital green innovation, but also offers new empirical evidence on how firms disclose and prioritize innovation under environmental and digital pressures, thereby generating important implications for both policymakers and enterprise managers.
Keywords: Digital green innovation; Mobile computing; Text mining; Annual reports; Statistical analysis

1 INTRODUCTION

Addressing climate change and advancing sustainable development have become urgent priorities on the global agenda, as reflected in initiatives such as the United Nations Sustainable Development Goals (SDGs) and the rapid expansion of ESG investment worldwide. In response, enterprises are facing increasing regulatory, market, and societal pressure to integrate green innovation into their core strategies. At the same time, the rapid diffusion of digital technologies—including artificial intelligence, big data analytics, the Internet of Things, and cloud computing—is reshaping production processes and business models. These technologies offer new opportunities for improving resource efficiency, enhancing environmental management, and supporting sustainable transformation in firms.

A growing body of literature has examined the relationship between digitalization and green innovation. Prior studies generally find that digital technologies can improve environmental performance and stimulate sustainable innovation within firms [1-4]. Empirical research in this area typically measures green innovation using structured data sources, such as patent counts, R&D expenditure, and financial indicators. For example, Petruzzelli et al. assessed green innovation through patent-based indicators and explored the technological and organizational factors influencing innovation outcomes [5]. Similarly, Yang and Zhu evaluated green innovation efficiency in China's manufacturing sector using structured input–output indicators and econometric methods such as the three-stage DEA and Malmquist index [6]. These approaches provide valuable insights into innovation performance and regional heterogeneity.

Despite these advances, existing measurement approaches remain heavily dependent on structured quantitative data. While such indicators capture observable innovation outputs and investments, they often overlook the contextual and strategic information embedded in corporate disclosures. Narrative documents such as annual reports frequently contain detailed descriptions of firms' digital transformation strategies, technological development, and environmental initiatives, yet they have received relatively limited attention in empirical assessments of green innovation. In addition, ESG-related datasets often suffer from inconsistencies and lack of standardization across data providers, which further constrains their reliability in evaluating firm-level sustainability performance [7].

These limitations point to two important research gaps. First, the heterogeneity of firm-level innovation strategies remains underexplored due to the lack of systematic analysis of large-scale textual data. Second, existing studies lack robust multidimensional frameworks that capture the interaction between digitalization and green innovation as reflected in firms' own disclosures.

To address these gaps, this study develops a data-driven framework to evaluate digital green innovation in Chinese manufacturing firms. Using a Python-based web crawler, we collect and analyze the 2024 annual reports of 1,285 A-share listed companies and apply text mining techniques to extract information on firms' digital transformation and green innovation activities. Building on recent literature, we construct a multidimensional indicator system consisting of eleven keywords related to digital technologies and green innovation. This framework enables a quantitative and comparable evaluation of firm-level innovation strategies using textual evidence from annual reports.

Our empirical analysis reveals substantial heterogeneity in digital green innovation across firms. In particular, enterprises undergoing digital transformation demonstrate stronger green innovation performance, while firms in heavily polluting industries exhibit higher levels of green innovation activity, likely reflecting stronger regulatory and policy pressures.

This study contributes to the literature in three main ways. First, it introduces a text-based measurement approach that complements conventional structured indicators in assessing green innovation. Second, it develops a multidimensional evaluation framework that integrates digital transformation with sustainability assessment. Third, it provides new firm-level evidence on the patterns and drivers of digital green innovation in China's manufacturing sector.

2 METHODS

Figure 1 provides an overview of the methodological framework employed in this study, which consists of four major phases:

·**Data Collection:** The process begins with the systematic collection of 2024 annual reports from 1,285 A-share listed manufacturing companies in China. A Python-based web crawler, compatible with both mobile and cloud platforms, is utilized to automate document retrieval and ensure data accuracy.

·**Keyword System Construction:** Based on a comprehensive review of recent academic literature and regulatory standards, a multidimensional indicator system is developed. Eleven core keywords are selected to capture diverse dimensions of digital and green innovation.

·**Data Processing and Analysis:** The unstructured annual report texts undergo several preprocessing steps, including tokenization, cleaning, OCR (optical character recognition), and standardization. Multivariate statistical analyses and machine learning methods—such as frequency distribution, clustering, box plots, radar charts, and heatmaps—are then applied to assess firm-level innovation capability and heterogeneity.

·**Results Interpretation and Visualization:** The analytical outputs are visualized to highlight distribution patterns, correlation structures, and typical case differences across enterprises. These insights enable the identification of digital green innovation leaders and provide practical guidance for both policymakers and corporate managers.

This integrated workflow supports automated, large-scale, and mobile-adapted data mining, and sets a robust foundation for dynamic empirical analysis in the context of digital green transformation.

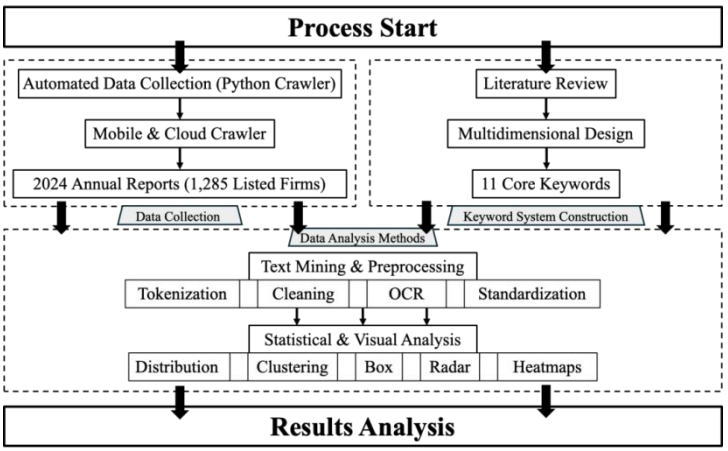

Figure 1 Research Workflow and Analytical Framework of This Study

2.1 Data Sources and Collection Process

This study utilizes the 2024 annual reports of 1,285 A-share listed manufacturing companies in China. Data were systematically collected from official disclosures by the China Securities Regulatory Commission (CSRC), Shenzhen and Shanghai Stock Exchanges, and major financial databases such as CNINFO and Eastmoney, with cross-verification to ensure accuracy and completeness.

A mobile-adapted, cloud-based Python crawler enabled flexible and real-time automated document retrieval. The pipeline integrated company list initialization, API construction, anti-crawling strategies, dynamic parameter management, JSON parsing, and automated filtering to remove irrelevant or duplicate files. Robust fingerprinting and manual checks guaranteed data integrity. Batch OCR, encoding standardization, regular-expression-based tokenization, and field validation were applied. Data from mobile devices were seamlessly integrated into cloud storage, with missing values supplemented from authoritative third-party sources. This automated workflow supports efficient, scalable, and reproducible unstructured text mining for downstream empirical analysis.

2.2 Design of Green Innovation Keywords

The multidimensional indicator system for green innovation was developed by synthesizing recent academic literature, regulatory standards, and expert consultation [8]. Eleven core keywords—directly relevant and text-identifiable—were

selected: “artificial intelligence,” “Internet,” “digital technology,” “green products,” “customer satisfaction,” “market share,” “production processes,” “performance,” “sales revenue,” “digital green,” and “number of patents.” The system builds on the frameworks of Yin et al. and Sartirana, and is tailored for the Chinese manufacturing context [9,10]. Each indicator dimension corresponds to quantifiable text features covering green product design, intelligent manufacturing, process innovation, market expansion, and patent output. The three main innovation perspectives—Technology, Market, and Management—are mapped into five evaluation dimensions (Table 1), capturing the breadth and depth of digital green innovation. Keyword statistics and normalization used Python-based tokenization and regular expressions, with results stored in a “firm–keyword–frequency” tensor. Reliability was validated by expert review and iterative testing, with refinement based on the latest standards. Lightweight algorithms enabled efficient, real-time analysis, making the system both robust and scalable for mobile deployment.

Table 1 Multi-dimensional Green Innovation Capability Evaluation System

Evaluation Dimension	Text Keywords
Digital Technology Empowerment	Artificial intelligence, Internet, Digital technology, Green products, Customer satisfaction
Market Expansion and Value Realization	Market share, Sales volume
Intelligent Manufacturing and Process Innovation	Production process, Performance
Innovation Output and Intellectual Property	Patent quantity, Digital green
Digital Marketing and Channel Innovation	Digital green, Customer satisfaction

2.3 Data Analysis Methods

To evaluate firm-level digital green innovation capability and explore its heterogeneity, this study employs a combination of descriptive statistics, correlation analysis, principal component analysis (PCA), and cluster analysis. These methods enable a comprehensive examination of the distribution, relationships, and structural patterns of innovation indicators.

2.3.1 Descriptive statistics and distribution analysis

Descriptive statistical methods are first used to summarize the overall distribution of green innovation indicators across firms. Box plots and frequency distributions are applied to visualize the dispersion, central tendency, and potential outliers in the innovation scores. These visualizations provide an intuitive understanding of the variability of digital green innovation performance among enterprises.

2.3.2 Correlation analysis

To examine the relationships among innovation indicators, Pearson and Spearman correlation coefficients are calculated. The Pearson correlation coefficient measures linear relationships between variables and is defined as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where x_i and y_i represent the observations of variables x and y, and $\bar{x}$ and $\bar{y}$ denote their respective means.

Spearman correlation coefficients are also computed to capture potential monotonic relationships among variables.

2.3.3 Principal Component Analysis (PCA)

Principal component analysis is employed to reduce dimensionality and identify the major underlying components of digital green innovation. PCA transforms a set of correlated variables into a smaller number of uncorrelated principal components through linear combinations of the original variables. The first principal component is obtained by maximizing the variance:

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \tag{2}$$

Where X_1,X_2,…,X_p are the standardized original variables and a_{1j} represents the loading coefficient. PCA helps identify the dominant innovation dimensions and evaluate the overall structure of the indicator system.

2.3.4 Cluster analysis

To explore heterogeneity among firms, K-means clustering is applied to group enterprises with similar innovation characteristics. The K-means algorithm partitions observations into K clusters by minimizing the within-cluster sum of squares (WCSS):

$$\min \sum_{k=1}^{K} \sum_{x_i \in C_k} |x_i - \mu_k|^2 \tag{3}$$

where $C_k$denotes cluster k and μ_k is the centroid of the cluster. This method allows firms to be categorized according to their digital green innovation profiles.

All statistical analyses are conducted using Python 3.11 with relevant libraries, including pandas, numpy, matplotlib, seaborn, and scikit-learn. Statistical inference is performed using a two-sided significance level of 0.05.

3 RESULTS AND DISCUSSION

Building on the integrated data acquisition and indicator system, this study conducts a multi-level empirical analysis of digital green innovation capabilities among 1,285 listed manufacturing firms in China, leveraging synchronized mobile and cloud platforms. This section examines sample characteristics, keyword distribution trends, heterogeneity, dimensional correlations, and representative enterprise cases, focusing on the unique insights enabled by large-scale text mining.

3.1 Sample Description and Statistical Characteristics

Table 2 summarizes the distribution of key innovation keywords at the enterprise level. Notable heterogeneity is observed: while some firms frequently mention “performance” and “sales revenue,” output-related terms such as “digital green” and “number of patents” remain sparse. This variation, captured through cloud-based data integration, highlights underlying differences in innovation disclosure and practice. Such patterns align with Pu , who identified pronounced imbalances and heterogeneity in digital green innovation across Chinese firms [11]. This text-driven approach reveals hierarchical differences that structured data alone often cannot detect, laying a robust foundation for subsequent heterogeneity and correlation analyses.

Table 2 Summary Statistics of Sample Firms and Main Variables

Indicator	Min	Max	Mean	Std. Dev.
Number of firms			1285	
Artificial intelligence frequency	0	41	1.65	3.44
Internet frequency	0	55	2.72	5.92
Digital technology frequency	0	25	0.81	2.06
Green product frequency	0	19	0.88	1.62
Customer product frequency	0	11	0.27	0.71
Market share frequency	0	29	2.18	3.45
Production process frequency	0	15	0.52	1.30
Performance frequency	0	85	9.68	10.93
Sales volume frequency	0	55	3.85	6.79
Digital green frequency	0	3	0.01	0.11
Patent quantity frequency	0	7	0.11	0.47

The results indicate clear stratification in digital and green innovation capabilities, driven by differences in digital transformation depth, green strategy, and policy orientation. This approach, leveraging unstructured text, allows for a more nuanced understanding of enterprise heterogeneity than traditional structured metrics.

3.2 Keyword Distribution and Overall Trends

Grouped bar charts (Figure 2) visualize the frequency distribution of innovation keywords. “Performance,” “sales revenue,” and “market share” dominate, suggesting firms’ primary focus on outputs and market expansion. Meanwhile, “artificial intelligence,” “Internet,” and “digital technology” show an upward trend, reflecting the growing synergy between digital infrastructure and green innovation—consistent with findings by Yang and Liu [12]. The low occurrence of output-oriented keywords (“digital green,” “number of patents”) suggests a gap between innovation input and realized outcomes, highlighting areas for policy and managerial improvement. The text-mining perspective here offers dynamic, real-time industry mapping that traditional indicators cannot provide.

3.3 Overall Distribution of Green Innovation Capability

Figure 3 presents box plots of green innovation capability scores across all firms. The distribution is notably dispersed: only a small subset of enterprises excel, while most remain at modest innovation levels. This pattern—low median, high upper quartile, numerous outliers—echoes the findings of Xie et al. on uneven green innovation in heavily polluting Chinese industries [13]. These disparities are shaped by internal resources, policy incentives, and technological accumulation. Importantly, the use of annual report texts enables more granular and timely detection of such patterns compared to conventional data sources, supporting stratified analysis and tailored policy design.

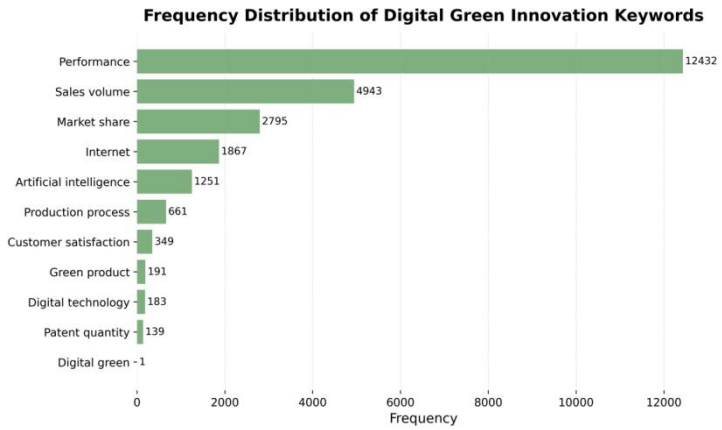

Figure 2 Total Frequency Distribution of Digital Green Innovation Keywords

3.4 Dimensional Correlation Analysis

The correlation heatmap (Figure 4) explores relationships among innovation dimensions. Output indicators ("performance," "market share," "sales revenue") are strongly correlated, while links between technical input keywords and outputs are weaker. This suggests a lag or inefficiency in translating digital investments into tangible innovation results—a point also observed by Wang et al. Real-time, mobile-enabled correlation analysis supports on-site diagnosis and more agile innovation management. These findings provide new evidence for optimizing innovation processes and highlight the distinct advantage of lightweight, AI-powered text analytics for industry-wide monitoring.

3.5 Case Enterprise Radar Chart Analysis

To further illustrate enterprise heterogeneity, radar charts (Figure 5) were generated for three top-performing firms, using both mobile data and real-time manager feedback. The analysis reveals substantial differences in dimensions such as "artificial intelligence," "market share," and "production processes," reflecting diverse innovation pathways. The radar chart feature, embedded in mobile platforms, enables managers to benchmark against industry leaders and quickly identify performance gaps. Compared with Eriandani and Winarno, these findings confirm the value of multidimensional, text-mining-based evaluation for industry segmentation and targeted policy support.

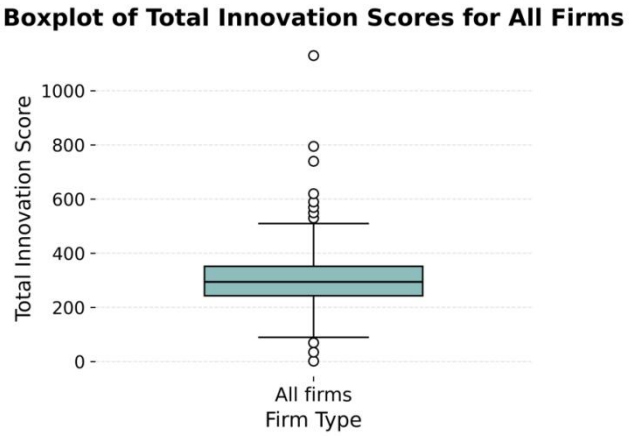

Figure 3 Boxplot of Total Innovation Scores for All Sampled Firms

3.6 Overall Synthesis and Key Findings

Synthesizing the above results, several key patterns and insights emerge regarding digital green innovation among Chinese manufacturing firms. First, the analysis reveals pronounced heterogeneity in innovation disclosure and capability at the firm level. While most enterprises concentrate on output-oriented dimensions such as "performance," "sales revenue," and "market share," only a small subset excel in digital and green innovation indicators, resulting in a highly skewed distribution of overall innovation scores. This pattern suggests the presence of leading innovators alongside a large base of lagging firms, reflecting industry-wide disparities shaped by internal resources, technological capabilities, and policy environments.

Second, the frequency and correlation analyses indicate that digital technology-related keywords ("artificial intelligence," "Internet," "digital technology") are on the rise, yet their transformation into tangible innovation outcomes remains limited, as evidenced by the weak linkage between technical input and output dimensions. This

finding highlights a persistent gap between digital investment and realized green innovation, underscoring the need for more effective strategies to bridge this divide.

Third, the radar chart analysis of representative enterprises further underscores the diversity of innovation pathways within the sector. Leading firms demonstrate strengths across multiple dimensions, while others exhibit highly specialized or unbalanced profiles. The use of real-time, mobile-enabled text mining and visualization not only facilitates timely benchmarking and diagnosis but also supports tailored policy and management interventions.

Collectively, these findings validate the effectiveness of large-scale, text-mining based approaches for capturing both explicit and implicit aspects of digital green innovation. The methodology provides a dynamic, scalable framework for industry-wide monitoring, enabling more granular, actionable insights than traditional structured data sources. Ultimately, this study provides new empirical evidence on digital green innovation in China's manufacturing sector.

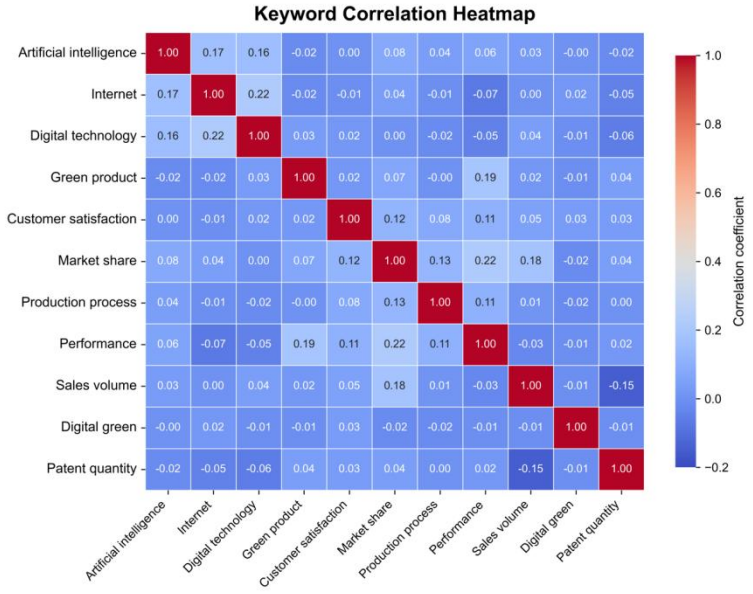

Figure 4 Heatmap of Correlation Coefficients Among Digital Green Innovation Keywords

3.7 Limitations and Future Directions

Despite its contributions, this study has several limitations. First, the granularity and disclosure norms of annual report texts may omit certain innovation activities. Second, keyword-based frequency analysis cannot fully capture deep semantic context or resolve ambiguity. Third, constraints on mobile device processing power limit advanced semantic extraction and model deployment. Finally, while mobile data collection and analysis were rigorously simulated, full-scale deployment in enterprise production settings awaits further validation. Future research should integrate advanced semantic models (e.g., BERT, GPT), extend multiyear tracking and multi-source data integration (e.g., patents, supply chain), and deepen field collaboration for real-world application. Further investigation is also warranted into the causal mechanisms linking innovation, policy, and market orientation.

Figure 5 Comparison of Green Innovation Capability Among Typical Firms, Visualized by Radar Charts

4 CONCLUSION

This study develops a data-driven framework to evaluate digital green innovation in Chinese manufacturing firms using large-scale textual analysis of corporate annual reports. Based on the 2024 annual reports of 1,285 A-share listed manufacturing companies, the study constructs a multidimensional indicator system consisting of eleven keywords

related to digital technologies and green innovation. By integrating text mining with multivariate statistical analysis, the framework enables a systematic assessment of firm-level innovation capability and heterogeneity.

The empirical results reveal several important findings. First, digital transformation and green innovation are increasingly intertwined within China's manufacturing sector, as reflected in the frequent occurrence of digitalization-related keywords such as "artificial intelligence," "Internet," and "digital technology." Second, substantial heterogeneity exists across firms in terms of digital green innovation capability. In particular, firms in heavily polluting industries tend to exhibit higher levels of green innovation activity, likely reflecting stronger regulatory pressures and policy incentives. Third, the relatively low frequency of output-oriented indicators such as "digital green" and "number of patents" suggests a potential gap between digital investment and the realization of tangible green innovation outcomes.

Despite these contributions, several limitations should be acknowledged. The keyword-based measurement approach may not fully capture the semantic complexity and strategic intent embedded in corporate disclosures. In addition, the current framework focuses primarily on textual information and does not yet integrate other important data sources such as patent quality, R&D expenditure, or market-based indicators. Future research could incorporate advanced natural language processing models and combine multi-source datasets to further improve the accuracy and robustness of digital green innovation measurement.

Overall, this study provides new empirical evidence on digital green innovation in China's manufacturing sector and offers a scalable methodological approach for analyzing firm-level innovation using large-scale textual data.

COMPETING INTERESTS

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

REFERENCES

[1] Ning J, Jiang X, Luo J. Relationship between enterprise digitalization and green innovation: A mediated moderation mode. Journal of Innovation & Knowledge, 2023, 8: 100326.

[2] Xu P, Chen L, Dai H. Pathways to Sustainable Development: Corporate Digital Transformation and Environmental Performance in China. Sustainability, 2023, 15(1): 256.

[3] He Z, Kuai L, Wang J. Driving mechanism model of enterprise green strategy evolution under digital technology empowerment: A case study based on Zhejiang enterprises. Business Strategy and the Environment, 2022: 1-22.

[4] Cheng H, Li Y, Pang Y, et al. Can digital transformation change a firm's green innovation strategy? Evidence from China's heavily polluting industries. Heliyon, 2024, 10: e24676.

[5] Petruzzelli A M, Dangelico R M, Rotolo D, et al. Organizational factors and technological features in the development of green innovations: Evidence from patent analysis. SSRN Electronic Journal, 2011.

[6] Yang H, Zhu X. Research on Green Innovation Performance of Manufacturing Industry and Its Improvement Path in China. Sustainability, 2022, 14(13): 8000.

[7] Kotsantonis S, Serafeim G. Four Things No One Will Tell You About ESG Data. SSRN Electronic Journal, 2019.

[8] Zhang G, Gao Y, Li G. Research on Digital Transformation and Green Technology Innovation: Evidence from China's Listed Manufacturing Enterprises. Sustainability, 2023, 15(8): 6425.

[9] Lin Z, Liang D, Li S. Environmental Regulation and Green Technology Innovation: Evidence from China's Heavily Polluting Companies. Sustainability, 2022, 14(19): 12180.

[10] Yin S, Yu Y. An adoption-implementation framework of digital green knowledge to improve the performance of digital green innovation practices for industry 5.0. Journal of Cleaner Production, 2022, 363: 132608.

[11] Yin S, Zhang N, Ullah K, et al. Enhancing Digital Innovation for the Sustainable Transformation of Manufacturing Industry: A Pressure-State-Response Framework. Systems, 2022, 10(3): 72.

[12] Sartirana C. Network Effects on Green Propensity and Determinants of Persistence in Green Innovation. Milano: University of Milano-Bicocca, 2024.

[13] Pu T. The Role of Digitalization in Enhancing Green Innovation Efficiency: Examination of Multiple Sub-Dimensions and Heterogeneity. SAGE Open, 2025.