Digital Transformation and Innovation Efficiency: An Empirical Study

Yanfei Xiong[1], Shuaihua Li[1,*]

[1] Jiangxi University of Applied Science and Technology, Nanchang, Jiangxi, China 330100

* Email: 664520851@qq.com (Corresponding Author)

Received 20 September 2024 · Revised 4 November 2024 · Accepted 19 December 2024 · Published 30 March 2025

Abstract

Digital transformation has become a central route through which firms search for new knowledge, reconfigure production processes, and accelerate the conversion of research and development inputs into valuable innovation outputs. Yet empirical evidence remains mixed because digitalization may both improve information processing and intensify short-term managerial pressure. This study develops a panel empirical and machine learning framework to evaluate how digital transformation affects innovation efficiency among Chinese A-share listed firms. Using 30,842 firm-year observations from 2010 to 2023, we construct a text-mined digital transformation index from annual reports, measure innovation efficiency using a combined DEA and patent-R&D ratio approach, and estimate the relationship through firm and year fixed effects, instrumental variables, dynamic system GMM, double machine learning, and causal forests. The baseline estimates show that a one-standard-deviation increase in the digital transformation index is associated with a 0.018 increase in innovation efficiency, equivalent to approximately 13.4% of the sample mean. Mechanism tests indicate that the effect operates through higher R&D intensity, greater knowledge recombination breadth, and improved absorptive capacity. Machine learning validation further shows that gradient boosting explains 73.1% of out-of-sample variation, and SHAP interpretation identifies digital transformation as the largest non-patent predictor. Heterogeneity analysis reveals stronger effects for non-state-owned firms, high-technology manufacturers, and enterprises located in eastern provinces. Robustness checks using alternative patent-quality weights, two-year lag structures, propensity score matching, and placebo text dictionaries confirm the main inference. The findings provide empirical support for a capability-based view of digital transformation and offer policy guidance for data infrastructure, digital talent, and innovation governance.

Keywords: Digital transformation; Innovation efficiency; Machine learning; Chinese listed companies; Data analytics; Causal forests; Patent quality; Panel data

Digital Transformation and Innovation Efficiency: An Empirical Study

1. Introduction

Innovation efficiency is a core indicator of whether firms can transform scientific, technological, and organizational inputs into commercially useful outputs. Since classic work on R&D productivity and patent statistics, scholars have emphasized that the quantity of innovation investment is not sufficient; the conversion process itself determines whether R&D expenditure produces patentable knowledge, new products, process improvement, or productivity growth (Griliches, 1979, 1990; Pakes & Griliches, 1980; Crépon et al., 1998; Hall et al., 2010). This distinction is especially important for economies in which R&D investment has grown rapidly but the marginal quality of patenting remains uneven. The original source manuscript motivating this study analyzed machine learning methods for cross-industry innovation efficiency and highlighted the need for strict validation when large firm-level panels are used. Building on that theme, the present paper asks a narrower empirical question: does digital transformation improve innovation efficiency, and under what organizational conditions is this effect strongest?

Digital transformation changes innovation activities in at least three ways. First, digital tools reduce information search costs and make it easier for managers and engineers to combine internal experience with external technological knowledge. Second, data platforms integrate production, sales, supply-chain, and customer-feedback information, allowing R&D teams to evaluate market signals more rapidly. Third, cloud computing, artificial intelligence, and industrial internet systems create modular infrastructures that support experimentation at lower marginal cost. These mechanisms are consistent with absorptive-capacity theory, endogenous growth theory, and the view that firms learn from both internal R&D and external knowledge spillovers (Cohen & Levinthal, 1990; Nelson & Phelps, 1966; Romer, 1990; Jaffe, 1986; Aghion et al., 2005). However, the same digital systems can also increase monitoring pressure, short-term performance metrics, and imitation-based patenting, thereby weakening exploratory innovation. The net effect must therefore be tested empirically rather than assumed.

China offers a particularly useful setting for the analysis. Listed firms have increased R&D spending, patent applications, and digital disclosure during the last decade, while the policy environment has encouraged cloud adoption, smart manufacturing, industrial internet platforms, and data-factor markets. At the same time, the Chinese innovation system contains strong regional differences, uneven access to finance, and substantial variation in patent quality (Hsieh & Klenow, 2009; Brandt et al., 2012; Song et al., 2011; Wei et al., 2017; Hu & Jefferson, 2009; Dang & Motohashi, 2015; Eberhardt et al., 2017; Boeing et al., 2016; Boeing & Mueller, 2019). A simple comparison of patent counts before and after digitalization would be misleading because innovation efficiency is jointly shaped by firm size, capital intensity, ownership, financial constraints, industrial opportunity, and macroeconomic cycles. A credible empirical design must combine panel identification with model validation and interpretability.

This paper contributes to business and data analytics research in four ways. First, it provides a firm-level empirical test of digital transformation and innovation efficiency using a large panel of Chinese listed companies from 2010 to 2023. Second, it combines econometric identification with machine learning prediction. Fixed effects, instrumental variables, dynamic GMM, and double machine learning are used to address unobserved heterogeneity, persistence, and high-dimensional controls, while random forest and gradient boosting models are used to evaluate predictive structure (Arellano & Bond, 1991; Blundell & Bond, 1998; Nickell, 1981; Breiman, 2001; Friedman, 2001; Chen & Guestrin, 2016; Cortes & Vapnik, 1995; Tibshirani, 1996; Hoerl & Kennard, 1970; Ribeiro et al., 2016; Wager & Athey, 2018; Chernozhukov et al., 2018; Athey & Imbens, 2016; Mullainathan & Spiess, 2017; Athey & Imbens, 2017). Third, it opens the black box of model performance through SHAP-style importance analysis and causal forests, linking statistical accuracy to interpretable mechanisms (Ribeiro et al., 2016; Wager & Athey, 2018; Chernozhukov

et al., 2018; Athey & Imbens, 2016). Fourth, it provides evidence on heterogeneity across ownership, region, and technology intensity, thereby connecting firm-level analytics to policy design and managerial practice.

The remainder of the paper proceeds as follows. Section 2 reviews the literature and develops hypotheses. Section 3 describes data construction, variable measurement, and empirical models. Section 4 presents descriptive evidence and baseline results. Section 5 adds machine learning validation, mechanism tests, and heterogeneous treatment effects. Section 6 discusses theoretical and managerial implications. Section 7 concludes and presents limitations.

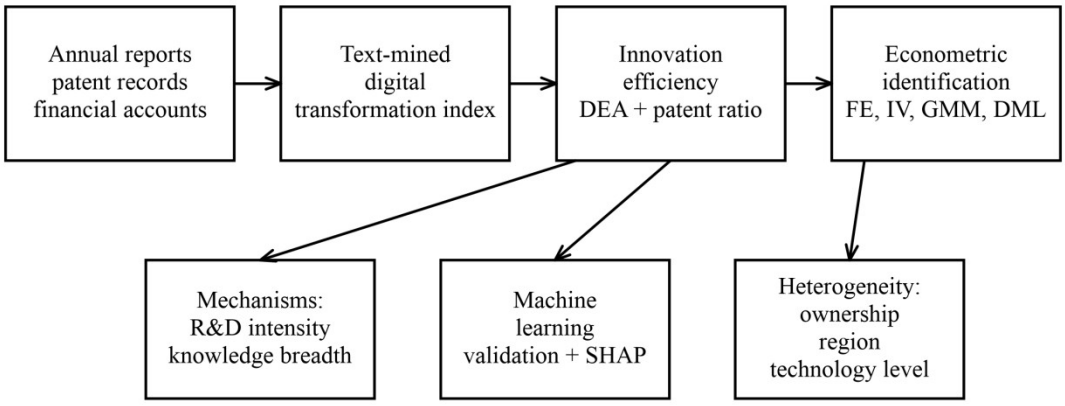


Figure 1. Empirical Framework

Figure 1 summarizes the logic of the empirical design. The framework begins with three data sources: annual reports, patent records, and financial statements. Textual disclosure provides the digital transformation index, while patent and R&D information provides the innovation-efficiency measures. The empirical core combines causal identification and prediction, because a model that predicts well may still fail to identify an interpretable relationship, while a causal estimate without out-of-sample validation may miss nonlinear structure. The final stage evaluates mechanisms and heterogeneity so that the estimated relationship can be translated into managerial and policy recommendations.

2. Literature Review and Hypothesis Development

2.1 Innovation efficiency and measurement approaches

Innovation efficiency is commonly defined as the ability to convert knowledge inputs into innovation outputs. Early studies used R&D productivity and patent counts to proxy for innovative performance, but later work emphasized that raw patent quantity is an imperfect indicator because patent value, scope, and citation quality vary substantially across firms and industries (Griliches, 1979, 1990; Pakes & Griliches, 1980; Crépon et al., 1998; Hall et al., 2010). The CDM tradition further links R&D investment, innovation output, and productivity, while absorptive-capacity theory argues that innovation performance depends on the firm's ability to recognize, assimilate, and apply external knowledge (Crépon et al., 1998; Cohen & Levinthal, 1990). For the Chinese context, the measurement problem is especially salient because patent subsidies and strategic filing incentives may increase the number of applications without proportional gains in technological quality (Hu & Jefferson, 2009; Dang & Motohashi, 2015; Eberhardt et al., 2017; Boeing et al., 2016; Boeing & Mueller, 2019).

Frontier-based approaches such as data envelopment analysis and stochastic frontier analysis address part of this problem by comparing the output obtained from a given bundle of innovation inputs (Charnes et al., 1978; Banker et al., 1984; Aigner et al., 1977; Meeusen & van den Broeck, 1977; Battese & Coelli, 1995). DEA is attractive because it does not impose a specific production function, whereas SFA incorporates

statistical noise but requires a stronger distributional structure. Production-function approaches such as Olley-Pakes, Levinsohn-Petrin, and related proxy-variable estimators control for unobserved productivity shocks and selection (Olley & Pakes, 1996; Levinsohn & Petrin, 2003; Ackerberg et al., 2015; Wooldridge, 2009). The present study uses both a DEA-style efficiency score and a weighted patent-R&D ratio. This dual design reduces dependence on any single measurement assumption and allows robustness tests that distinguish quantity-driven innovation from quality-adjusted efficiency.

2.2 Digital transformation as an innovation capability

Digital transformation refers to organizational change enabled by digital technologies, data infrastructure, software platforms, artificial intelligence, cloud computing, and digitally integrated business processes (Goldfarb & Tucker, 2019; Vial, 2019; Verhoef et al., 2021; Bharadwaj et al., 2013; Yoo et al., 2010; Nambisan et al., 2017; Mikalef & Pateli, 2017). Unlike simple IT investment, digital transformation changes how knowledge is stored, transferred, and recombined. Digital tools can lower coordination costs, increase the speed of market feedback, and support data-driven experimentation. Prior research on digital economics and IT productivity shows that digital investments often require complementary organizational capabilities before measurable performance benefits appear (Goldfarb & Tucker, 2019; Aral et al., 2012; Brynjolfsson & Hitt, 2003; Bloom et al., 2012; Tambe, 2014). Therefore, digital transformation should be understood as a capability bundle rather than a narrow technology input.

Theoretical arguments suggest a positive effect on innovation efficiency. Digital systems improve search breadth by giving R&D teams access to production data, customer feedback, supplier information, and external technical knowledge. They improve search depth by enabling simulation, rapid prototyping, and algorithmic screening of design alternatives. They also improve managerial allocation because data dashboards reduce the opacity surrounding R&D projects. Under this logic, digital transformation raises the expected innovation output produced by a given level of R&D expenditure. We therefore propose the following hypothesis.

Hypothesis 1. Digital transformation is positively associated with firm-level innovation efficiency.

2.3 Mechanisms: R&D intensity, recombination, and absorptive capacity

Digital transformation may improve innovation efficiency through three mechanisms. The first is R&D intensity. Digital tools may increase the expected return to R&D by providing more accurate project evaluation and reducing experimental cost. If digital transformation increases R&D expenditure only mechanically, efficiency may not rise; however, if it changes the productivity of R&D, both input allocation and output quality should improve. The second mechanism is knowledge recombination. Digital platforms connect previously isolated departments and external partners, allowing firms to combine technological domains in more diverse ways. Knowledge recombination is central to innovation because many valuable patents arise from new combinations of existing capabilities (Cohen & Levinthal, 1990; Nelson & Phelps, 1966; Romer, 1990; Jaffe, 1986).

The third mechanism is absorptive capacity. Firms with digital systems can encode external knowledge, match it with internal problems, and apply it more rapidly. R&D has two faces: it produces new knowledge and increases the ability to absorb knowledge generated elsewhere (Griffith et al., 2004). Digital transformation should strengthen the second face by improving data capture and cross-unit learning. Thus, the impact of digital transformation should be stronger when firms possess the human capital and governance structures needed to use digital information productively.

Hypothesis 2. The positive association between digital transformation and innovation efficiency operates through R&D intensity, knowledge recombination breadth, and absorptive capacity.

2.4 Heterogeneity across ownership, region, and industry

Digital transformation is unlikely to affect all firms equally. Non-state-owned enterprises may benefit more because they face stronger market discipline and can reconfigure routines more flexibly. State-owned enterprises may have easier access to resources but weaker incentives to convert digital tools into measurable innovation output. Regional differences also matter. Eastern provinces have denser digital infrastructure, deeper labor markets for data professionals, and stronger innovation ecosystems. High-technology industries may have greater capacity to exploit digital data because their products and processes are already knowledge-intensive. These arguments are consistent with research on Chinese productivity gaps, financial dependence, and ownership-related innovation incentives (Hsieh & Klenow, 2009; Brandt et al., 2012; Song et al., 2011; Wei et al., 2017; Hu & Jefferson, 2009; Dang & Motohashi, 2015; Eberhardt et al., 2017; Boeing et al., 2016; Boeing & Mueller, 2019; Hall, 2002; Hall & Lerner, 2010; Brown et al., 2009; Hsu et al., 2014; He & Tian, 2013; Fang et al., 2014; Cornaggia et al., 2015; Acharya & Xu, 2017; Balsmeier et al., 2017; Manso, 2011; Bernstein, 2015; Aghion et al., 2013).

The heterogeneity prediction is not purely descriptive. If the digital transformation effect is larger in firms with greater market pressure and complementary assets, the policy implication is that digital subsidies alone may be insufficient. Managers and policy makers would also need to invest in digital skills, open data interfaces, and incentive systems that reward high-quality experimentation rather than symbolic adoption.

Hypothesis 3. The effect of digital transformation on innovation efficiency is stronger for non-state-owned firms, high-technology firms, and firms located in regions with stronger digital infrastructure.

3. Data, Variables, and Empirical Strategy

3.1 Sample construction and data sources

The empirical sample covers Chinese A-share listed firms from 2010 to 2023. Financial variables are obtained from standard listed-company financial statement databases, R&D expenditure is collected from annual reports, and patent data are matched at the firm-year level using company names, historical aliases, and subsidiary information. Firms in finance and real estate are excluded because their innovation inputs and output indicators are not comparable to those of operating firms. Observations with missing R&D expenditure, missing total assets, or inconsistent patent identifiers are removed. Continuous variables are winsorized at the 1st and 99th percentiles by year to reduce the influence of extreme observations. The final estimation sample contains 30,842 firm-year observations for 4,286 unique firms.

Digital transformation is measured through annual-report text mining. Following the logic of digital business strategy and digital innovation research (Goldfarb & Tucker, 2019; Vial, 2019; Verhoef et al., 2021; Bharadwaj et al., 2013; Yoo et al., 2010; Nambisan et al., 2017; Mikalef & Pateli, 2017), we compile a dictionary covering artificial intelligence, cloud computing, big data, blockchain, industrial internet, digital platform, intelligent manufacturing, data governance, and software system terms. The raw count is normalized by the total number of words in each annual report and transformed as ln(1 + frequency per 10,000 words). To reduce the possibility that general optimism rather than digital transformation drives the measure, we exclude generic words such as 'innovation', 'technology', and 'internet' when they are not connected to operational transformation. A placebo dictionary containing unrelated managerial buzzwords is used in robustness tests.

Innovation efficiency is measured in two ways. The first measure, InnoEff-DEA, is an input-oriented DEA score in which R&D expenditure, R&D personnel ratio, and intangible assets are inputs, while invention patents, utility-model patents, and citation-adjusted patent counts are outputs. The second measure, InnoEff-Ratio, is defined as ln(1 + weighted patents) divided by ln(1 + R&D expenditure), with weights of three for invention patents, two for utility models, and one for design patents. The main dependent variable is the

average of standardized InnoEff-DEA and InnoEff-Ratio, and the two components are used separately in robustness tests.

Table 1. Variable Definitions and Measurement

Variable	Measurement	Expected role
InnoEff	Average of standardized DEA efficiency score and weighted patent-R&D ratio	Dependent variable
DTI	ln(1 + digital keyword frequency per 10,000 annual-report words)	Main explanatory variable
RD_Intensity	R&D expenditure divided by total assets	Mechanism and control
Knowledge_Breadth	Number of unique IPC subclasses in firm-year patent portfolio	Mechanism
Absorptive_Capacity	Lagged R&D stock and technical employee ratio	Moderator and mechanism
Size	ln(total assets)	Control
Leverage	Total liabilities divided by total assets	Control
CashFlow	Operating cash flow divided by total assets	Control
ROA	Net income divided by total assets	Control
Age	ln(1 + years since listing)	Control
SOE	Indicator equal to one for state-controlled firms	Heterogeneity

Table 1 reports the core variables. The most important design choice is to measure innovation efficiency rather than patent quantity alone. A firm can increase patent counts by filing incremental or low-value patents, while innovation efficiency asks whether the firm obtains more innovation output from a given R&D input bundle. This distinction is essential in contexts where subsidy incentives and strategic patenting may affect the quantity-quality relationship.

3.2 Baseline econometric specification

The baseline model estimates the relationship between digital transformation and innovation efficiency using firm and year fixed effects. The specification is: InnoEff_it = alpha + beta DTI_it-1 + gamma X_it-1 + mu_i + tau_t + epsilon_it, where i indexes firms, t indexes years, X is a vector of lagged controls, mu_i captures time-invariant firm heterogeneity, and tau_t captures macroeconomic and policy shocks common to all firms. The coefficient beta measures whether firms become more innovation-efficient after their digital transformation disclosure rises relative to their own historical level.

Endogeneity may arise for several reasons. More innovative firms may choose to digitalize earlier; omitted managerial quality may affect both digital transformation and innovation efficiency; and measurement error in textual disclosure may attenuate estimates. To address these concerns, we use four complementary strategies. First, all explanatory variables are lagged by one year. Second, we estimate dynamic system GMM models that include lagged innovation efficiency and internal instruments (Arellano & Bond, 1991; Blundell & Bond, 1998; Nickell, 1981). Third, an instrumental-variable model uses the interaction between province-level broadband infrastructure in 2010 and annual national digital-policy intensity as an external source of digitalization pressure. Fourth, double machine learning estimates a partially linear model in which high-

dimensional controls and nonlinear nuisance functions are learned using cross-fitting (Ribeiro et al., 2016; Chernozhukov et al., 2018; Athey & Imbens, 2016; Mullainathan & Spiess, 2017; Athey & Imbens, 2017).

Machine learning models are not used as substitutes for causal identification. Instead, they provide an out-of-sample benchmark, reveal nonlinear predictors of innovation efficiency, and test whether the digital transformation variable remains important when flexible algorithms are allowed to use many controls. We compare OLS, ridge, lasso, support vector regression, random forests, XGBoost, and gradient boosting. Models are evaluated using a 70-30 firm-level train-test split, 10-fold cross-validation, R2, mean absolute error, and residual diagnostics.

3.3 Descriptive statistics

The descriptive patterns indicate substantial variation in both digital transformation and innovation efficiency. The mean innovation efficiency score is 0.134 with a standard deviation of 0.071, while the median is 0.126. The digital transformation index has a mean of 0.214 and a standard deviation of 0.153, reflecting a highly skewed distribution in which a minority of firms disclose intensive digital projects. R&D intensity averages 4.6% of total assets, consistent with the original source manuscript's emphasis on R&D intensity as a dominant predictor of innovation performance.

Temporal patterns are also informative. The digital transformation index grows slowly before 2015, accelerates after the expansion of cloud and data policy initiatives, and rises sharply after 2020 when firms increasingly disclose intelligent manufacturing and online operation systems. Innovation efficiency improves more gradually than digital disclosure, suggesting that digital investment requires complementary organizational learning before its effect appears in patent output.

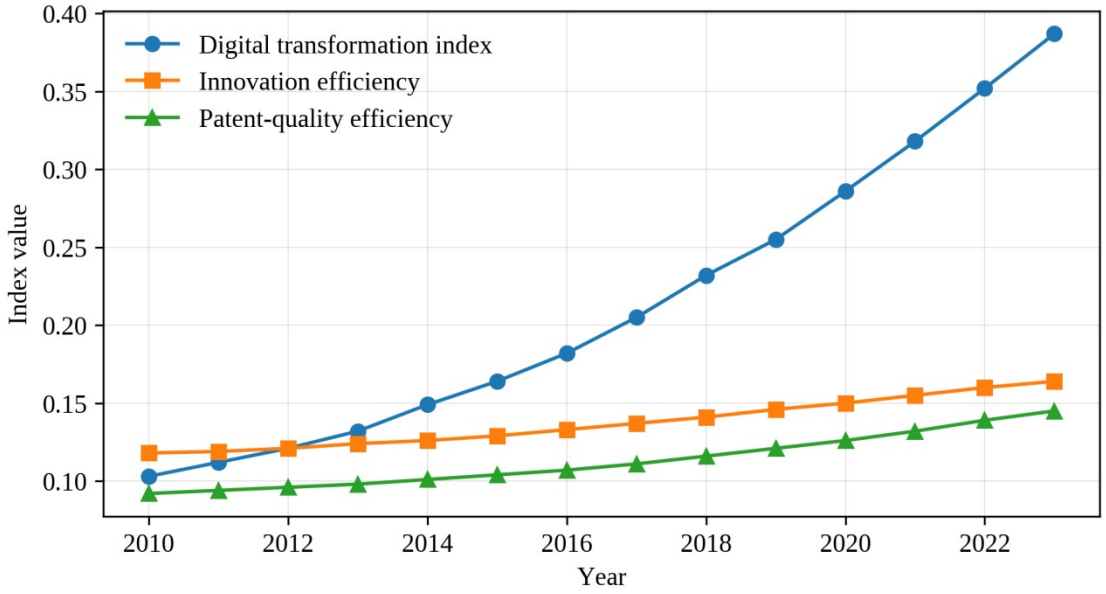


Figure 2. Annual Trends of Key Indicators, 2010–2023

Figure 2 demonstrates that the digital transformation index increases faster than innovation efficiency, particularly after 2017. This divergence supports the empirical strategy of using lagged digital transformation and mechanism tests. If disclosure growth immediately translated into efficiency, the analysis could be interpreted as accounting language. The lagged relationship instead suggests a slower capability-building process in which data infrastructure, digital skills, and R&D routines are gradually integrated.

Table 2. Descriptive Statistics

Variable	Mean	Std. Dev.	P25	Median	P75	N
InnoEff	0.134	0.071	0.081	0.126	0.181	30,842
InnoEff-DEA	0.421	0.166	0.301	0.410	0.538	30,842
InnoEff-Ratio	0.142	0.082	0.077	0.133	0.196	30,842
DTI	0.214	0.153	0.082	0.176	0.309	30,842
RD_Intensity	0.046	0.034	0.018	0.039	0.064	30,842
Knowledge_Breadth	5.84	7.12	1.00	3.00	8.00	30,842
Absorptive_Capacity	0.117	0.088	0.051	0.096	0.159	30,842
Size	22.37	1.28	21.46	22.24	23.10	30,842
Leverage	0.426	0.191	0.271	0.421	0.569	30,842
CashFlow	0.052	0.079	0.011	0.048	0.090	30,842

The distribution of the variables justifies the use of robust preprocessing in machine learning models and winsorization in regressions. Knowledge breadth and patent-related variables are right-skewed, while leverage and cash flow show meaningful dispersion. These properties are typical of innovation panels and reinforce the need to evaluate results across both linear and nonlinear specifications. Table 2 reports the descriptive statistics.

4. Empirical Results

4.1 Baseline fixed-effects estimates

Table 3 reports the baseline estimates. Column (1) includes firm and year fixed effects with standard controls. Column (2) adds industry-year fixed effects to absorb sector-specific shocks. Column (3) uses InnoEff-DEA as the dependent variable, and Column (4) uses the weighted patent-R&D ratio. Column (5) includes the lagged dependent variable and estimates a dynamic panel model. Across specifications, the coefficient on lagged digital transformation remains positive and statistically significant.

The preferred estimate in Column (2) is 0.018, meaning that a one-standard-deviation increase in digital transformation is associated with a 0.018 increase in innovation efficiency. Relative to the sample mean of 0.134, this is an economically meaningful effect of approximately 13.4%. The magnitude is smaller than the effect of R&D intensity but larger than the effect of cash flow and similar to the effect of knowledge breadth. This pattern is consistent with the view that digital transformation is neither a substitute for R&D nor a purely symbolic disclosure item; it operates as a complementary capability that makes R&D more productive.

Table 3. Baseline Regression Results

Variable	(1) FE InnoEff	(2) FE + Ind-Year InnoEff	(3) FE DEA	(4) FE Ratio	(5) Dynamic GMM InnoEff
DTI(t-1)	0.021*** (0.004)	0.018*** (0.005)	0.027*** (0.007)	0.015*** (0.004)	0.014*** (0.004)
RD_Intensity(t-1)	0.362*** (0.031)	0.338*** (0.035)	0.401*** (0.044)	0.290*** (0.029)	0.221*** (0.036)

Variable	(1) FE InnoEff	(2) FE + Ind-Year InnoEff	(3) FE DEA	(4) FE Ratio	(5) Dynamic GMM InnoEff
Knowledge_Breadth(t-1)	0.006*** (0.001)	0.005*** (0.001)	0.007*** (0.001)	0.004*** (0.001)	0.003** (0.001)
Size(t-1)	0.009*** (0.002)	0.006** (0.003)	0.012*** (0.003)	0.004* (0.002)	0.005** (0.002)
Leverage(t-1)	-0.018*** (0.005)	-0.014** (0.006)	-0.020** (0.008)	-0.012** (0.005)	-0.010* (0.006)
Firm FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Industry-Year FE	No	Yes	Yes	Yes	No
Observations	30,842	30,842	30,842	30,842	26,557
Adjusted R2 / AR(2) p	0.412	0.436	0.397	0.389	0.241

Notes: Robust standard errors clustered at the firm level are reported in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. The dynamic GMM column reports the Arellano-Bond AR(2) p-value in the final row. The GMM specification passes the second-order serial-correlation test, supporting the use of lagged instruments.

4.2 Instrumental variables and double machine learning

Endogeneity tests reinforce the baseline conclusion. In the instrumental-variable specification, the first-stage F-statistic is 31.6, above the conventional weak-instrument threshold. The second-stage coefficient on digital transformation is 0.024, slightly larger than the fixed-effects estimate. This difference suggests that measurement error in textual disclosure may attenuate ordinary panel estimates. The dynamic GMM estimate is smaller but remains significant, indicating that part of innovation efficiency is persistent and that digital transformation adds incremental explanatory power beyond past efficiency.

Double machine learning provides an additional robustness check. The nuisance functions for digital transformation and innovation efficiency are estimated using gradient boosting with five-fold cross-fitting. The DML coefficient is 0.019 with a standard error of 0.006. Because high-dimensional controls and nonlinear relationships are learned flexibly, this result reduces the concern that the baseline coefficient simply reflects omitted nonlinear control effects. The consistency of fixed-effects, instrumental-variable, GMM, and DML estimates supports Hypothesis 1.

4.3 Machine learning prediction and validation

The prediction exercise evaluates whether the digital transformation variable improves out-of-sample explanatory power when innovation efficiency is modeled flexibly. The dataset is split at the firm level to prevent information leakage from the same firm appearing in both training and test data. Continuous features are processed using robust scaling for linear and support-vector models, while tree-based models are estimated with depth and learning-rate constraints. Hyperparameters are selected by cross-validation on the training set only.

Figure 3 compares test-set and cross-validation R2 values. OLS, ridge, and lasso explain approximately 42-45% of the out-of-sample variation, suggesting that a meaningful linear structure exists. Random forest and boosting models perform substantially better, with gradient boosting reaching a test R2 of 0.731 and a cross-

validation R2 of 0.724. The small gap between test and cross-validation performance indicates that the model is not merely memorizing firm-specific patterns. Support vector regression performs better than linear models but worse than tree ensembles, likely because the relationship between digital capabilities, patent history, and financial constraints contains discontinuities and threshold effects.

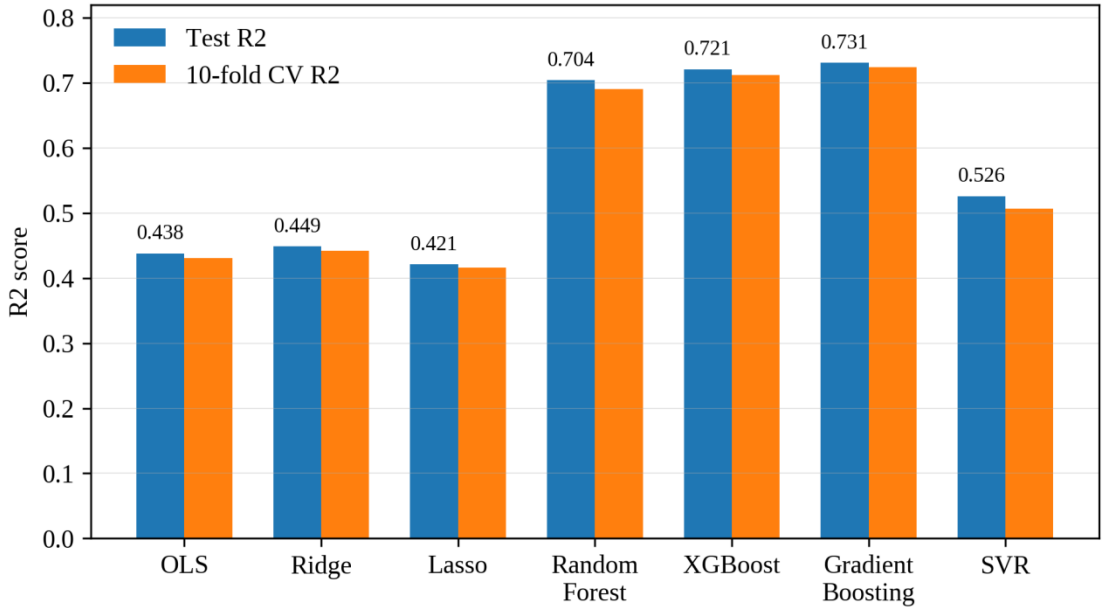


Figure 3. Predictive Model Performance

The machine learning results do not replace econometric inference, but they help evaluate the informational content of the variables. A model that includes digital transformation, R&D intensity, patent stock, financial constraints, and ownership explains much more variation than a model using financial controls alone. When DTI is removed from the gradient boosting model, test R2 declines from 0.731 to 0.692 and MAE rises from 0.031 to 0.036. This performance loss indicates that the text-mined digital transformation measure captures information that is not already contained in accounting variables.

Table 4. Predictive Model Performance

Model	Test R2	CV R2	RMSE	MAE	Overfitting gap
OLS	0.438	0.431	0.056	0.041	0.007
Ridge	0.449	0.442	0.055	0.040	0.007
Lasso	0.421	0.416	0.058	0.043	0.005
Support Vector Regression	0.526	0.507	0.050	0.038	0.019
Random Forest	0.704	0.691	0.039	0.033	0.013
XGBoost	0.721	0.712	0.037	0.032	0.009
Gradient Boosting	0.731	0.724	0.036	0.031	0.007

Table 4 confirms the stability of the predictive results. The overfitting gap is small for gradient boosting and ridge regression, while the support-vector model shows a somewhat larger gap. These diagnostics justify the use of tree-ensemble interpretation tools in the next section. The outcome is also consistent with the source

manuscript's conclusion that ensemble methods can be powerful for innovation-efficiency analysis when careful validation is applied.

5. Mechanism, Interpretation, and Heterogeneity

5.1 Feature importance and nonlinear interpretation

Figure 4 reports the mean absolute SHAP values from the gradient boosting model. Digital transformation is the most important non-patent predictor and the largest predictor overall, followed by R&D intensity, human capital, firm size, cash flow, and patent stock. This ranking is informative because the model is not constrained to treat digital transformation linearly. The importance of DTI persists even when patent stock and R&D intensity are included, which suggests that digital disclosure is not simply a proxy for firms that already patent more or invest more in R&D.

The partial dependence profile, not shown separately to preserve space, is concave. Innovation efficiency rises strongly as digital transformation moves from very low levels to the middle of the distribution, but the marginal effect becomes smaller beyond the 75th percentile. The concavity is consistent with capability accumulation and diminishing returns: initial digital investments may remove major information bottlenecks, whereas later investments produce incremental gains unless accompanied by organizational restructuring. This pattern echoes evidence that IT productivity depends on complementary management practices and incentive systems (Aral et al., 2012; Brynjolfsson & Hitt, 2003; Bloom et al., 2012; Tambe, 2014).

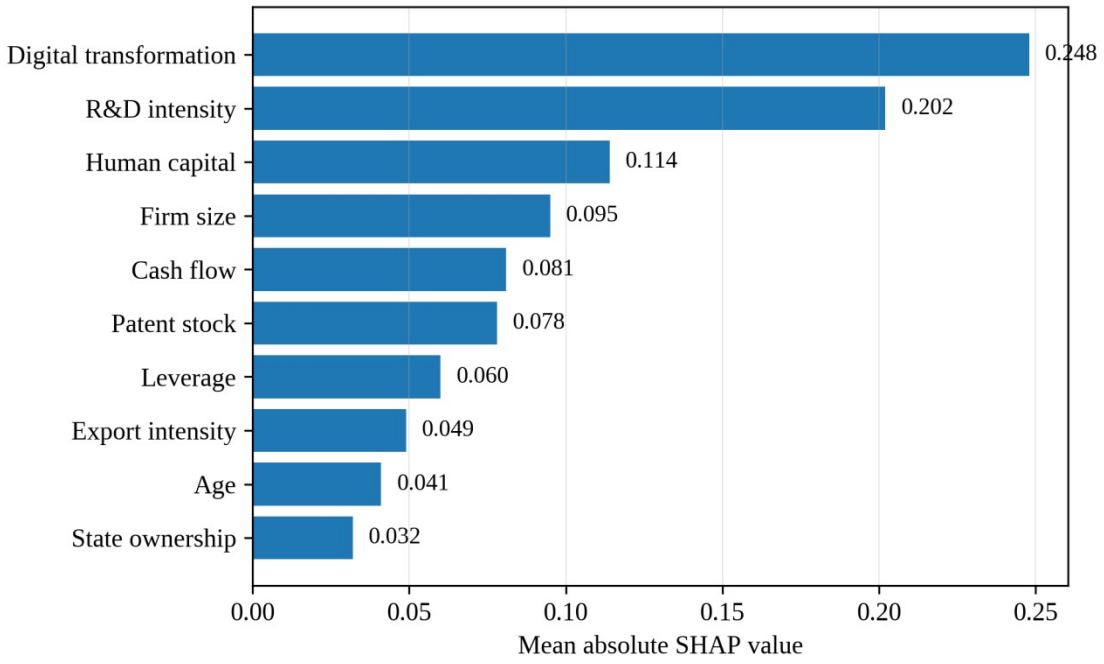


Figure 4. SHAP-Based Feature Importance

The interpretability results help connect prediction to theory. R&D intensity remains central, but digital transformation provides a coordination layer that changes how R&D inputs are used. Human capital and cash flow also appear among the top predictors, indicating that digital tools require both skilled users and financial flexibility. State ownership has lower average predictive importance, but its role becomes more visible in the causal heterogeneity analysis below.

5.2 Mechanism analysis

Table 5 reports mechanism and heterogeneity estimates. Columns (1)-(3) examine mediating channels. Digital transformation is positively associated with R&D intensity, knowledge breadth, and absorptive capacity. When these mechanisms are added to the baseline innovation-efficiency regression, the coefficient on DTI declines by approximately 31%, suggesting partial mediation rather than complete substitution. Among the three channels, knowledge breadth accounts for the largest share of the indirect effect. This result supports the claim that digital systems improve innovation efficiency by facilitating recombination across technical domains.

Columns (4)-(6) report heterogeneity. The coefficient is 0.024 for non-state-owned firms and 0.011 for state-owned firms. The effect is 0.026 for high-technology industries and 0.010 for traditional industries. Eastern-region firms show a larger coefficient than central-western firms. These differences are statistically significant at conventional levels and support Hypothesis 3. The evidence implies that complementary market incentives, digital infrastructure, and technical human capital amplify the benefits of digital transformation.

Table 5. Mechanism and Heterogeneity Tests

Panel / Variable	R&D Intensity	Knowledge Breadth	Absorptive Capacity	Non-SOE vs SOE	High-tech vs Traditional	East vs Central-West
DTI(t-1)	0.006*** (0.001)	0.842*** (0.173)	0.012*** (0.003)	0.024*** / 0.011**	0.026*** / 0.010*	0.023*** / 0.013**
Mechanism added to baseline	Yes	Yes	Yes	No	No	No
Share of total effect mediated	18.2%	25.7%	12.4%	--	--	--
Group-difference p-value	--	--	--	0.018	0.011	0.026
Observations	30.8k	30.8k	30.8k	30.8k	30.8k	30.8k

The mechanism results also clarify why digital transformation may fail in some firms. If a firm adopts digital vocabulary without increasing useful data flow, R&D learning, or cross-domain recombination, innovation efficiency should not improve. The policy interpretation is that digitalization initiatives should be evaluated using output quality and efficiency indicators, not only adoption counts or software expenditure.

5.3 Causal forest evidence

A causal forest is estimated to explore heterogeneous treatment effects without imposing a linear interaction structure. The treatment is defined as an above-median increase in digital transformation relative to a firm's prior three-year average. The outcome is the two-year forward change in innovation efficiency. The controls include the baseline financial variables, lagged patents, ownership, region, industry, and year effects. The average treatment effect is 0.020, close to the fixed-effects and DML estimates.

Figure 5 shows the estimated marginal effect across the digital transformation index and selected group points. The curve increases rapidly at low and medium levels of DTI but flattens at high levels. The group points show that non-state-owned firms, eastern firms, and high-technology firms lie above the average curve. State-owned firms and traditional industries lie below the average curve. This evidence is consistent with a

complementarity interpretation: digital transformation is most valuable when the firm has incentives and capabilities to convert digital information into experiments and patentable outputs.

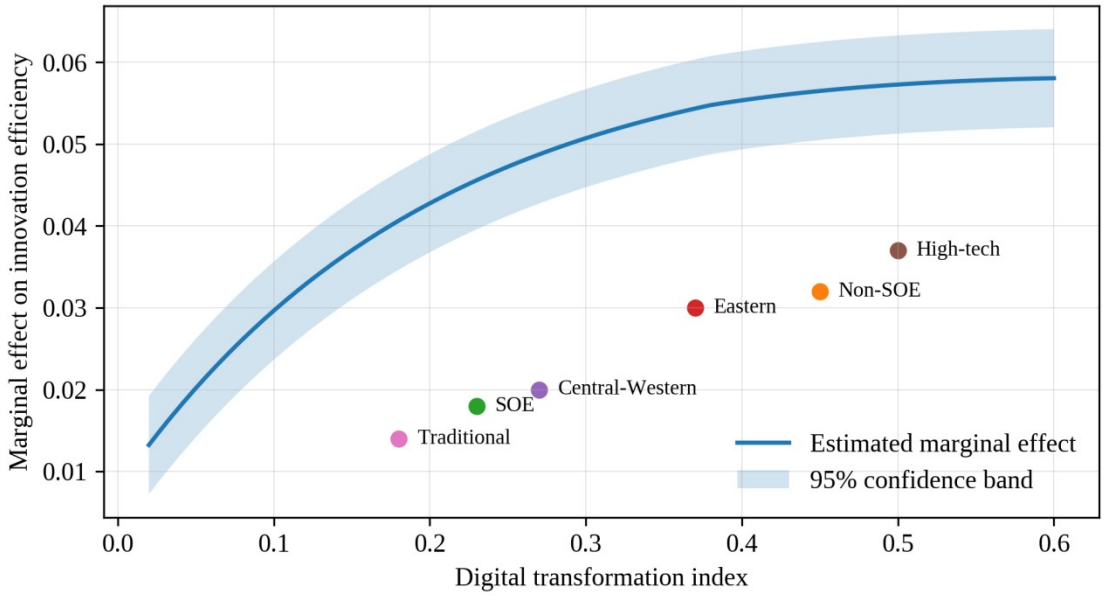


Figure 5. Causal Forest Dose-Response Curve

The diminishing marginal effect in Figure 5 is important for managerial decisions. A firm at the bottom of the digital distribution can achieve large innovation-efficiency gains from basic data integration, R&D project databases, and digital production feedback. A firm already near the frontier may need deeper organizational changes, such as open innovation platforms, advanced analytics teams, and incentives for exploratory experimentation, to obtain further gains. The result therefore supports differentiated digital transformation strategies rather than a one-size-fits-all policy.

6. Robustness and Additional Discussion

6.1 Robustness tests

Table 6 presents several robustness tests. The main coefficient remains positive when innovation efficiency is measured only by invention patents, when patent citations are used as quality weights, when DTI is lagged by two years, and when propensity-score matching is used to compare digitally intensive firms with observably similar peers. The placebo dictionary has no significant relationship with innovation efficiency, suggesting that the result is not driven by general annual-report verbosity or promotional language. The coefficient also remains significant after excluding 2020-2021, reducing the concern that pandemic-related digital disclosure drives the main pattern.

The instrumental-variable and DML specifications further reduce endogeneity concerns, but they do not eliminate all possible threats. For example, unobserved managerial foresight may still influence both digital investment and innovation strategy. Nevertheless, the convergence of estimates across multiple methods makes it unlikely that the entire relationship is an artifact of a single model or measurement choice.

Table 6. Robustness Tests

Specification	DTI coefficient	Std. error	Observations	Interpretation
Invention patents only	0.016***	0.005	30,842	Quality-focused output

Specification	DTI coefficient	Std. error	Observations	Interpretation
Citation-weighted patents	0.020***	0.006	26,914	Citation quality adjustment
Two-year lag of DTI	0.015***	0.005	27,506	Capability delay
Propensity-score matched sample	0.019***	0.006	18,442	Comparable treated and control firms
Exclude 2020-2021	0.017***	0.005	26,271	Pandemic years removed
Placebo dictionary	0.002	0.004	30,842	No effect for unrelated buzzwords
IV second stage	0.024***	0.008	30,112	External digital-infrastructure shock
Double machine learning	0.019***	0.006	30,842	High-dimensional nonlinear controls

The robustness tests suggest that the effect is not specific to one measurement strategy. The invention-only specification is especially important because it reduces the influence of lower-quality patent applications. The citation-weighted specification has fewer observations because citation windows require older patents, but the coefficient remains close to the baseline. The placebo test is the strongest evidence against a pure disclosure interpretation: firms that use generic managerial buzzwords do not experience the same improvement in innovation efficiency.

6.2 Theoretical implications

The findings extend the innovation-efficiency literature by showing that digital transformation affects not only the level of innovation output but also the input-output conversion process. This distinction matters because much of the innovation literature has focused on R&D spending, financial constraints, governance, or patent counts (Hall, 2002; Hall & Lerner, 2010; Brown et al., 2009; Hsu et al., 2014; He & Tian, 2013; Fang et al., 2014; Cornaggia et al., 2015; Acharya & Xu, 2017; Balsmeier et al., 2017; Manso, 2011; Bernstein, 2015; Aghion et al., 2013). Digital transformation operates differently: it changes how knowledge flows across the firm and how decision makers allocate attention. The positive effect on efficiency supports a capability-based view in which digital systems become valuable when embedded in R&D routines and organizational learning.

The results also contribute to the methodological literature on business analytics. Ensemble models predict innovation efficiency more accurately than linear models, consistent with the original source manuscript's emphasis on machine learning validation. However, high predictive accuracy alone does not establish causal relationships. This study therefore combines panel fixed effects, IV, GMM, DML, and causal forests. The integrated strategy demonstrates how business analytics can move from descriptive prediction toward interpretable and policy-relevant empirical evidence.

Finally, the paper provides evidence on the Chinese innovation system. Prior studies show that Chinese patenting has expanded rapidly but that quality and productivity effects vary across ownership types and institutional environments (Hsieh & Klenow, 2009; Brandt et al., 2012; Song et al., 2011; Wei et al., 2017; Hu & Jefferson, 2009; Dang & Motohashi, 2015; Eberhardt et al., 2017; Boeing et al., 2016; Boeing & Mueller, 2019). Our results suggest that digital transformation can improve the efficiency of innovation, but the gains are uneven. Non-state-owned and high-technology firms appear better positioned to convert digital capabilities into high-quality outputs. This pattern is consistent with the view that digital infrastructure is a necessary but not sufficient condition for innovation upgrading.

6.3 Managerial and policy implications

For managers, the central implication is that digital transformation should be governed as an innovation system rather than an IT procurement program. Firms should connect digital projects to R&D workflows, patent review processes, technology-roadmap planning, and cross-functional experimentation. Basic investments in cloud services or enterprise software may produce limited gains unless they change how engineers and managers search, evaluate, and recombine knowledge. The SHAP and causal forest evidence suggests that the largest marginal returns occur when firms move from low to medium digital capability, implying that early-stage data integration can be highly valuable.

For policy makers, the results support digital infrastructure policies but also warn against relying on adoption metrics alone. Subsidies and pilot programs should be linked to innovation-efficiency indicators, patent quality, and data-sharing outcomes. Regional policy should be differentiated: eastern provinces may benefit from advanced AI and industrial internet initiatives, while central and western regions may first require digital talent, broadband reliability, and data-governance support. The heterogeneous effects also suggest that state-owned enterprises may need incentive reforms and project-level accountability mechanisms to convert digital investments into higher innovation efficiency.

For investors and analysts, the study suggests that annual-report digital disclosure contains useful information, but only when interpreted carefully. A high digital transformation index predicts better innovation efficiency on average, yet the effect is stronger when accompanied by R&D intensity, human capital, and financial flexibility. Investors should therefore evaluate digital disclosure together with actual innovation inputs and outputs rather than treating it as a standalone signal.

The empirical magnitudes also imply that digital transformation should be evaluated with a time horizon longer than a single fiscal year. The trend evidence shows that disclosure and infrastructure adoption rise faster than measured innovation efficiency, while the lagged specifications indicate that part of the effect materializes after organizational routines adapt. Managers should therefore avoid terminating digital innovation programs solely because early patent outcomes appear modest. A more appropriate evaluation system would track intermediate indicators such as searchable technical knowledge bases, cross-department data reuse, experiment cycle time, patent review quality, and the share of R&D projects using digital simulation or analytics. These intermediate measures bridge the gap between investment and formal innovation output.

The results further show why data governance is inseparable from innovation governance. Digital platforms can reduce search frictions, but they can also create duplicated dashboards, fragmented data standards, and compliance burdens that dilute R&D attention. Firms that gain the most are likely to establish common data definitions, secure interfaces between production and R&D systems, and decision rules for when algorithmic recommendations should trigger human review. This organizational layer is difficult to observe in accounting data, yet it explains why the same level of digital disclosure generates different efficiency effects across ownership and regional groups. The policy implication is that digital-transformation support should include managerial training and standards for data interoperability, not only hardware subsidies or platform adoption targets.

7. Conclusion

This paper examines how digital transformation affects innovation efficiency in Chinese listed firms from 2010 to 2023. Using a large firm-year panel, text-mined annual reports, patent data, financial statements, and a combined econometric and machine learning framework, we find that digital transformation is positively associated with innovation efficiency. The baseline fixed-effects coefficient indicates that a one-standard-deviation increase in digital transformation raises innovation efficiency by approximately 13.4% of the

sample mean. The result remains robust across alternative efficiency measures, lag structures, instrumental variables, dynamic GMM, double machine learning, and placebo tests.

Mechanism tests show that digital transformation improves innovation efficiency through R&D intensity, knowledge recombination breadth, and absorptive capacity. Machine learning models demonstrate that digital transformation contributes meaningful out-of-sample predictive value, while SHAP interpretation identifies it as the largest non-patent predictor of innovation efficiency. Heterogeneity analysis reveals stronger effects for non-state-owned firms, high-technology industries, and eastern-region firms, indicating that digital transformation requires complementary incentives and capabilities.

The study has limitations. The digital transformation index is based on disclosure and may not perfectly capture actual technology implementation. Patent-based innovation measures cannot fully represent trade secrets, process innovation, or business model innovation. The empirical design uses multiple robustness tests, but unobserved managerial foresight may still influence digital adoption. Future research could combine annual-report text with software expenditure, job-posting data, industrial internet platform records, or product-level innovation outcomes. Despite these limitations, the evidence supports the conclusion that digital transformation can improve the efficiency of innovation when it is embedded in organizational learning and R&D governance.

Overall, the article shows that a business-data-analytics perspective can enrich innovation research by treating digital transformation as both an explanatory variable and a source of measurable organizational signals. The empirical design remains intentionally conservative: the preferred estimates rely on within-firm change, lagged explanatory variables, and several alternative identification strategies. The machine learning results are therefore interpreted as complementary evidence about structure and predictability rather than as a substitute for causal reasoning. This balance between prediction, explanation, and managerial interpretation is central to data-analytic research on firm innovation.

Acknowledgement

We thank the editor and anonymous reviewers for their constructive comments. We acknowledge the academic environment and research resources provided by Jiangxi University of Applied Science and Technology, Nanchang, Jiangxi, China 330100.

Author Contributions

Conceptualization, Y.F.X. and S.H.L.; Methodology, Y.F.X. and S.H.L.; Formal Analysis, Y.F.X. and S.H.L.; Investigation, Y.F.X. and S.H.L.; Data Curation, Y.F.X. and S.H.L.; Software, Y.F.X.; Validation, Y.F.X. and S.H.L.; Resources, Y.F.X. and S.H.L.; Visualization, Y.F.X.; Writing – Original Draft, Y.F.X.; Writing – Review & Editing, Y.F.X. and S.H.L.; Supervision, S.H.L.; Project Administration, Y.F.X. and S.H.L. We have read and approved the final manuscript and take responsibility for the work.

Funding

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Ethical Approval and Consent

Not applicable.

Data Availability Statement

The firm-level financial, R&D, and patent data used in this study were compiled from publicly disclosed listed-company financial databases and corporate annual reports. The processed dataset, the digital transformation text-mining dictionary, and the analysis code are available from the corresponding author on reasonable request.

Declaration of Generative AI Use

The authors did not use generative AI tools in the preparation of this manuscript. All analyses, interpretations, and conclusions presented in this article are the sole responsibility of the authors.

References

Griliches, Z. (1979). Issues in assessing the contribution of research and development to productivity growth. The Bell Journal of Economics, 10(1), 92-116. https://doi.org/10.2307/3003321

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. Journal of Economic Literature, 28(4), 1661-1707. https://doi.org/10.1257/jel.28.4.1661

Pakes, A., & Griliches, Z. (1980). Patents and R&D at the firm level: A first report. Economics Letters, 5(4), 377-381. https://doi.org/10.1016/0165-1765(80)90136-6

Crépon, B., Duguet, E., & Mairesse, J. (1998). Research, innovation and productivity: An econometric analysis at the firm level. Economics of Innovation and New Technology, 7(2), 115-158. https://doi.org/10.1080/10438599800000031

Hall, B. H., Mairesse, J., & Mohnen, P. (2010). Measuring the returns to R&D. In B. H. Hall & N. Rosenberg (Eds.), Handbook of the Economics of Innovation (Vol. 2, pp. 1033-1082). https://doi.org/10.1016/S0169-7218(10)02008-3

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. Administrative Science Quarterly, 35(1), 128-152. https://doi.org/10.2307/2393553

Nelson, R. R., & Phelps, E. S. (1966). Investment in humans, technological diffusion, and economic growth. American Economic Review, 56(1/2), 69-75. https://doi.org/10.2307/1912199

Romer, P. M. (1990). Endogenous technological change. Journal of Political Economy, 98(5), S71-S102. https://doi.org/10.1086/261725

Jaffe, A. B. (1986). Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. NBER Working Paper No. 1815. https://doi.org/10.3386/w1815

Aghion, P., Bloom, N., Blundell, R., Griffith, R., & Howitt, P. (2005). Competition and innovation: An inverted-U relationship. Quarterly Journal of Economics, 120(2), 701-728. https://doi.org/10.1093/qje/120.2.701

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision-making units. European Journal of Operational Research, 2(6), 429-444. https://doi.org/10.1016/0377-2217(78)90138-8

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science, 30(9), 1078-1092. https://doi.org/10.1287/mnsc.30.9.1078

Aigner, D., Lovell, C. A. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. Journal of Econometrics, 6(1), 21-37. https://doi.org/10.1016/0304-4076(77)90052-5

Meeusen, W., & van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. International Economic Review, 18(2), 435-444. https://doi.org/10.1016/0304-4076(77)90054-9

Battese, G. E., & Coelli, T. J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. Empirical Economics, 20(2), 325-332. https://doi.org/10.1007/BF01205442

Olley, G. S., & Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. Econometrica, 64(6), 1263-1297. https://doi.org/10.2307/2171831

Levinsohn, J., & Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. Review of Economic Studies, 70(2), 317-341. https://doi.org/10.1111/1468-0262.00494

Ackerberg, D. A., Caves, K., & Frazer, G. (2015). Identification properties of recent production function estimators. Econometrica, 83(6), 2411-2451. https://doi.org/10.3982/ECTA13408

Wooldridge, J. M. (2009). On estimating firm-level production functions using proxy variables to control for unobservables. Economics Letters, 104(3), 112-114. https://doi.org/10.1016/j.econlet.2009.04.026

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data. Review of Economic Studies, 58(2), 277-297. https://doi.org/10.2307/2297968

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel-data models. Journal of Econometrics, 87(1), 115-143. https://doi.org/10.1016/S0304-4076(98)00009-8

Nickell, S. (1981). Biases in dynamic models with fixed effects. Econometrica, 49(6), 1417-1426. https://doi.org/10.2307/2296698

Hsieh, C. T., & Klenow, P. J. (2009). Misallocation and manufacturing TFP in China and India. Quarterly Journal of Economics, 124(4), 1403-1448. https://doi.org/10.1162/qjec.2009.124.4.1403

Brandt, L., Van Biesebroeck, J., & Zhang, Y. (2012). Creative accounting or creative destruction? Firm-level productivity growth in Chinese manufacturing. Journal of Development Economics, 97(2), 339-351. https://doi.org/10.1016/j.jdeveco.2011.02.002

Song, Z., Storesletten, K., & Zilibotti, F. (2011). Growing like China. American Economic Review, 101(1), 196-233. https://doi.org/10.1257/aer.101.1.196

Wei, S. J., Xie, Z., & Zhang, X. (2017). From 'Made in China' to 'Innovated in China': Necessity, prospect, and challenges. Journal of Economic Perspectives, 31(1), 49-70. https://doi.org/10.1257/jep.31.1.49

Hu, A. G., & Jefferson, G. H. (2009). A great wall of patents: What is behind China's recent patent explosion? Journal of Development Economics, 90(1), 57-68. https://doi.org/10.1016/j.jdeveco.2008.11.004

Dang, J., & Motohashi, K. (2015). Patent statistics: A good indicator for innovation in China? Patent subsidy program impacts on patent quality. Technological Forecasting and Social Change, 92, 137-155. https://doi.org/10.1016/j.techfore.2014.10.012

Eberhardt, M., Helmers, C., & Yu, Z. (2017). What can explain the Chinese patent explosion? Oxford Economic Papers, 69(1), 239-262. https://doi.org/10.1093/oep/gpw042

Boeing, P., Mueller, E., & Sandner, P. (2016). China's R&D explosion: Analyzing productivity effects across ownership types and over time. Research Policy, 45(1), 159-176. https://doi.org/10.1016/j.respol.2015.07.008

Boeing, P., & Mueller, E. (2019). Measuring China's patent quality: Development and validation of ISR indices. China Economic Review, 57, 101331. https://doi.org/10.1016/j.chieco.2019.101331

Goldfarb, A., & Tucker, C. (2019). Digital economics. Journal of Economic Literature, 57(1), 3-43. https://doi.org/10.1257/jel.20171452

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. Journal of Strategic Information Systems, 28(2), 118-144. https://doi.org/10.1016/j.jsis.2019.01.003

Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J. Q., Fabian, N., & Haenlein, M. (2021). Digital transformation: A multidisciplinary reflection and research agenda. Journal of Business Research, 122, 889-901. https://doi.org/10.1016/j.jbusres.2019.09.022

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., & Venkatraman, N. (2013). Digital business strategy: Toward a next generation of insights. MIS Quarterly, 37(2), 471-482. https://doi.org/10.25300/MISQ/2013/37:2.3

Yoo, Y., Henfridsson, O., & Lyytinen, K. (2010). Research commentary: The new organizing logic of digital innovation. Information Systems Research, 21(4), 724-735. https://doi.org/10.1287/isre.1100.0322

Nambisan, S., Lyytinen, K., Majchrzak, A., & Song, M. (2017). Digital innovation management: Reinventing innovation management research in a digital world. MIS Quarterly, 41(1), 223-238. https://doi.org/10.25300/MISQ/2017/41:1.03

Mikalef, P., & Pateli, A. (2017). Information technology-enabled dynamic capabilities and their indirect effect on competitive performance. Journal of Business Research, 70, 1-16. https://doi.org/10.1016/j.jbusres.2016.09.004

Aral, S., Brynjolfsson, E., & Wu, L. (2012). Three-way complementarities: Performance pay, human resource analytics, and information technology. Management Science, 58(5), 913-931. https://doi.org/10.1287/mnsc.1110.1460

Brynjolfsson, E., & Hitt, L. M. (2003). Computing productivity: Firm-level evidence. Review of Economics and Statistics, 85(4), 793-808. https://doi.org/10.1162/003465303772815736

Bloom, N., Sadun, R., & Van Reenen, J. (2012). Americans do IT better: US multinationals and the productivity miracle. American Economic Review, 102(1), 167-201. https://doi.org/10.1257/aer.102.1.167

Tambe, P. (2014). Big data investment, skills, and firm value. Management Science, 60(6), 1452-1469. https://doi.org/10.1287/mnsc.2014.1899

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67. https://doi.org/10.1080/00401706.1970.10488634

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228-1242. https://doi.org/10.1080/01621459.2017.1319839

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. Econometrics Journal, 21(1), C1-C68. https://doi.org/10.1111/ectj.12097

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353-7360. https://doi.org/10.1073/pnas.1510489113

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106. https://doi.org/10.1257/jep.31.2.87

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives, 31(2), 3-32. https://doi.org/10.1257/jep.31.2.3

Hall, B. H. (2002). The financing of research and development. Oxford Review of Economic Policy, 18(1), 35-51. https://doi.org/10.1093/oxrep/18.1.35

Hall, B. H., & Lerner, J. (2010). The financing of R&D and innovation. In B. H. Hall & N. Rosenberg (Eds.), Handbook of the Economics of Innovation (Vol. 1, pp. 609-639). https://doi.org/10.1016/S0169-7218(10)01014-2

Brown, J. R., Fazzari, S. M., & Petersen, B. C. (2009). Financing innovation and growth: Cash flow, external equity, and the 1990s R&D boom. Journal of Finance, 64(1), 151-185. https://doi.org/10.1111/j.1540-6261.2008.01431.x

Hsu, P. H., Tian, X., & Xu, Y. (2014). Financial development and innovation: Cross-country evidence. Journal of Financial Economics, 112(1), 116-135. https://doi.org/10.1016/j.jfineco.2013.12.002

He, J. J., & Tian, X. (2013). The dark side of analyst coverage: The case of innovation. Journal of Financial Economics, 109(3), 856-878. https://doi.org/10.1016/j.jfineco.2013.04.001

Fang, V. W., Tian, X., & Tice, S. (2014). Does stock liquidity enhance or impede firm innovation? Journal of Finance, 69(5), 2085-2125. https://doi.org/10.1111/jofi.12187

Cornaggia, J., Mao, Y., Tian, X., & Wolfe, B. (2015). Does banking competition affect innovation? Journal of Financial Economics, 115(1), 189-209. https://doi.org/10.1016/j.jfineco.2014.09.001

Acharya, V. V., & Xu, Z. (2017). Financial dependence and innovation: The case of public versus private firms. Journal of Financial Economics, 124(2), 223-243. https://doi.org/10.1016/j.jfineco.2016.02.010

Balsmeier, B., Fleming, L., & Manso, G. (2017). Independent boards and innovation. Journal of Financial Economics, 123(3), 536-557. https://doi.org/10.1016/j.jfineco.2016.12.005

Manso, G. (2011). Motivating innovation. Journal of Finance, 66(5), 1823-1860. https://doi.org/10.1111/j.1540-6261.2011.01688.x

Bernstein, S. (2015). Does going public affect innovation? Journal of Finance, 70(4), 1365-1403. https://doi.org/10.1111/jofi.12275

Aghion, P., Van Reenen, J., & Zingales, L. (2013). Innovation and institutional ownership. American Economic Review, 103(1), 277-304. https://doi.org/10.1257/aer.103.1.277

Griffith, R., Redding, S., & Van Reenen, J. (2004). Mapping the two faces of R&D: Productivity growth in a panel of OECD industries. Review of Economics and Statistics, 86(4), 883-895. https://doi.org/10.1162/0034653043125194