ISSN:2790-1688 Volume-14-(2025)

# The Impact of Technology Investment on Innovation Outcomes: A Quantitative Study Based on Patent Data

Meiqi Lu

The University of Hong Kong;

lumeiqilouise@163.com

**Abstract.** Innovation serves as a cornerstone of economic development and global competitiveness, with technology investment playing a pivotal role in driving inventive progress. This study investigates the intricate relationship between technology investment and innovation outcomes, utilizing patent data as a quantifiable proxy for innovation. By integrating comprehensive publicly available datasets with advanced machine learning methodologies, this research analyzes how financial and human capital investments shape both the quantity and quality of patents. Key findings reveal the significant influence of factors such as R&D expenditure and firm size, alongside the presence of non-linear dynamics and interaction effects. These insights not only bridge critical gaps in existing literature but also provide actionable recommendations for firms and policymakers seeking to optimize technology investment strategies and foster sustainable innovation ecosystems.

**Keywords:** technology investment; innovation outcomes; patent data; machine learning; R&D expenditure; sustainable innovation.

## 1. Introduction

Innovation is widely acknowledged as a cornerstone of economic growth and global competitiveness, serving as a catalyst for industrial advancement and societal progress. It facilitates the creation of new products, processes, and business models, enhancing productivity and addressing global challenges such as sustainability and health. At the heart of innovation lies technology investment, particularly in research and development (R&D), which forms the bedrock for scientific discovery, technological breakthroughs, and long-term economic benefits. Patents, as formal records of inventive activity, offer an established metric for evaluating both the scale and impact of innovation efforts. Despite the clear importance of these investments, understanding how they translate into measurable innovation outcomes remains an intricate and multifaceted challenge.

A significant body of literature has explored the link between technology investment and innovation. Early studies established a positive correlation between R&D expenditure and innovation outputs, measured through patent activity. However, these investigations often employ linear models that oversimplify the relationship, failing to capture the underlying complexity of innovation processes. Innovation is inherently non-linear, characterized by dynamic feedback loops, cross-disciplinary interactions, and varying returns on investment across different contexts. Moreover, traditional models frequently neglect critical contextual factors such as firm size, industry dynamics, and regional variations, which significantly influence the effectiveness of technology investments.

The complexities of innovation processes also extend to the temporal dimension, as the outcomes of R&D investments often exhibit significant time lags. For instance, the commercialization of patented technologies may take years, influenced by market conditions, regulatory frameworks, and organizational capabilities. Furthermore, heterogeneity among firms, including differences in absorptive capacity, managerial expertise, and strategic priorities—compounds the challenge of accurately modeling the relationship between technology investment and innovation.

This study seeks to address these limitations by adopting advanced analytical approaches that go beyond traditional econometric methods. Machine learning techniques, with their capacity to model complex, non-linear relationships and uncover hidden patterns in large datasets, offer a powerful

alternative for analyzing the multifaceted dynamics of innovation. These methods enable a more comprehensive examination of how financial and human capital investments interact with contextual variables to influence patent-related outcomes.

To better understand the multifaceted relationship between technology investment and innovation, this study seeks to answer the following questions:

1. How do financial and human capital investments in technology influence patent-related innovation outcomes, particularly in terms of quantity and quality?
2. What are the non-linear patterns and interaction effects between technology investment variables and contextual factors?
3. How can machine learning techniques improve modeling and prediction of innovation outcomes compared to traditional econometric methods?

Building upon these research questions, the key objectives of this study are:

1. Quantitatively analyze the relationship between technology investment and innovation outcomes, using patent data to measure both the quantity and quality of innovation.
2. Apply machine learning methodologies to identify and model non-linear dynamics, interaction effects, and contextual influences on innovation outcomes.
3. Develop actionable insights and strategic recommendations for firms and policymakers to optimize their technology investment strategies, enhancing innovation capacity.

By integrating machine learning methodologies with robust patent and financial datasets, this study aims to advance the methodological toolkit for innovative research. It contributes to the literature by addressing critical gaps, such as the limited exploration of non-linear dynamics and the underrepresentation of contextual influences. Beyond its academic contributions, the research has practical implications for decision-makers, offering evidence-based strategies to enhance innovation capacity in an era of rapid technological change and heightened global competition.

## 2. Literature Review

### 2.1 Technology Investment and Innovation

Technology investment encompasses the financial and human resources allocated to develop or acquire technological capabilities. Among these, research and development (R&D) investment is widely regarded as a primary driver of innovation, with firms dedicating substantial budgets to proprietary technology development (Griliches, 1990). However, R&D alone is often insufficient; external knowledge acquisition—such as through mergers and acquisitions (M&As)—also serves as a crucial strategy, allowing firms to integrate novel technologies and accelerate innovation (Arora et al., 2016). Additionally, human capital, particularly skilled labor in science and engineering, plays a pivotal role in translating R&D efforts into tangible outputs, as firms with stronger absorptive capacity can more effectively leverage external knowledge (Cohen & Levinthal, 1990).

The relationship between technology investment and innovation has been extensively examined in the literature. Early studies established a foundational link between R&D expenditure and patent activity, with patents widely accepted as proxies for innovation output (Griliches, 1990). Schmookler (1966) further emphasized the role of economic demand in shaping technological advancements, arguing that financial commitments to R&D must be strategically aligned with market needs to drive competitiveness. While patent counts are commonly used to measure innovation output (Meyer, 2000), this approach has limitations—it fails to distinguish between incremental and groundbreaking inventions. To address this, later studies such as Hall et al. (2005) expanded the scope by analyzing patent citations, which serve as a better indicator of an innovation's technological and economic impact. Additional contributions by Mansfield (1980) and Jaffe (1986) reinforced the significance of R&D spillovers, highlighting how technological advancements diffuse across firms and industries.

Collectively, these studies underscore that technology investment is not merely a financial decision but a strategic imperative for sustaining innovation-driven growth. However, traditional

methods of measuring innovation outputs often fall short in capturing the complex and dynamic nature of the innovation process, necessitating more advanced analytical approaches.

### 2.2 Absorptive Capacity and Firm-Level Dynamics

Cohen and Levinthal (1990) introduced the concept of absorptive capacity, defining it as a firm's ability to recognize, assimilate, and exploit external knowledge. Their seminal work demonstrated that higher R&D investments not only increase patent output but also enhance firms' ability to integrate emerging technologies effectively. Expanding on this, Aghion et al. (1992) investigated innovation in competitive environments, showing that firms with greater absorptive capacity achieve higher returns on R&D.

Zahra and George (2002) refined the absorptive capacity framework by differentiating between potential absorptive capacity (the ability to acquire and understand knowledge) and realized absorptive capacity (the ability to apply knowledge in innovation processes). More recent research highlights the role of organizational structures, managerial practices, and inter-organizational networks in maximizing absorptive capacity (Lane et al., 2006; Volberda et al., 2010). Firm-specific factors—including size, structure, strategic orientation, and industry characteristics—significantly influence how effectively technology investments translate into innovation.

This body of work suggests that firm-level heterogeneity plays a crucial role in shaping the innovation process, yet many empirical studies still overlook contextual variations when analyzing R&D productivity. Addressing these limitations requires methodologies capable of capturing firm-specific dynamics and non-linear relationships in innovation processes.

### 2.3 Traditional Econometric Approaches and Research Gaps

Early empirical research primarily relied on linear econometric models to establish relationships between technology investment and innovation outcomes. Production function models, for instance, treated patents as a direct output of R&D expenditure, assuming a constant rate of return on investment (Griliches, 1990). While these models laid the foundation for empirical innovation research, they often oversimplified the innovation process, neglecting critical feedback loops and diminishing returns to R&D investment. To refine innovation measurement, subsequent research incorporated patent citations to assess an innovation's technological significance and market impact (Hall et al., 2005). However, even these improved models faced significant limitations, including:

1. Over-reliance on linear models, which fail to capture the non-linear dynamics and interaction effects that characterize innovation processes (Griliches, 1990; Hall et al., 2005).
2. Omission of unobserved heterogeneity, such as organizational culture, managerial expertise, and serendipitous discoveries, which significantly influence innovation outcomes but are difficult to quantify.
3. Static treatment of time lags, as most studies relied on cross-sectional data, overlooking the delayed effects of R&D on patent generation and commercialization (Griliches, 1990).
4. Limited exploration of industry- and firm-level heterogeneity, particularly in high-tech and knowledge-intensive sectors, where innovation patterns differ significantly (Cohen & Levinthal, 1990; Zahra & George, 2002).

Moreover, despite extensive research on R&D investments and patent-based innovation measures, existing studies have underutilized advanced computational techniques, particularly machine learning, which can uncover hidden patterns, complex dependencies, and non-linear relationships in innovation processes. Machine learning approaches offer a promising alternative to traditional econometric models by addressing these limitations and enabling a more data-driven, adaptive, and predictive understanding of innovation dynamics. This study seeks to bridge these gaps by integrating econometric analysis with machine learning techniques, providing a more comprehensive, high-resolution perspective on the factors driving innovation outcomes. The findings aim to generate practical insights for firms and policymakers, offering evidence-based strategies to optimize technology investment and enhance innovation capacity.

### 2.4 Machine Learning in Innovation Analysis

While traditional econometric models have provided valuable insights into linear relationships, they often struggle to capture the non-linear interactions, contextual heterogeneities, and dynamic feedback loops inherent in the innovation process. Innovation is influenced by multiple interdependent factors, including financial and human capital investments, firm-specific characteristics, and external market conditions, which interact in complex ways that traditional regression-based models fail to fully capture. Recent advances in machine learning (ML) have demonstrated their potential as powerful analytical tools that can model these complexities, uncover hidden patterns, and enhance predictive accuracy in innovation studies. Unlike conventional models that rely on predefined functional forms, ML techniques learn from data, enabling greater flexibility and adaptability in identifying intricate dependencies among technology investment variables and innovation outcomes. By leveraging large datasets and adaptive algorithms, ML approaches offer a more comprehensive understanding of how R&D expenditures, absorptive capacity, and industry-specific factors contribute to innovation performance.

Various ML methodologies have been applied in innovation research, each offering distinct advantages. Supervised learning models, such as Random Forest, Gradient Boosting, and Support Vector Machines, are particularly effective in predicting innovation outcomes based on structured variables like R&D spending, firm characteristics, and policy interventions. For instance, Cockburn et al. (2018) applied ML techniques to analyze AI-related patents, identifying clusters of technological advancements and their economic impacts. Meanwhile, unsupervised learning approaches, including k-means clustering, topic modeling, and word embeddings, have been used for pattern recognition in large, unstructured datasets, such as patent texts, scientific publications, and corporate disclosures. Chen et al. (2021) demonstrated the effectiveness of deep learning models in evaluating patent quality by analyzing the textual features of patent abstracts, providing more nuanced assessments of innovation significance beyond simple patent counts. Furthermore, deep learning architectures, such as Neural Networks, Transformer Models, and Convolutional Neural Networks, have enabled more granular analyses of semantic relationships within patent descriptions and research articles, offering new ways to assess the evolution of technological knowledge. Additionally, causal machine learning techniques, including Causal Forests, Bayesian Networks, and Double Machine Learning, have been explored to infer cause-and-effect relationships between R&D investments and innovation outcomes. Athey and Imbens (2017) highlighted the potential of ML in enhancing causal inference, allowing for a more precise understanding of the factors that drive technological progress.

Despite these advancements, the application of machine learning in innovation research faces several challenges. One major concern is interpretability, as many ML models, particularly deep learning methods, function as black-box algorithms, making it difficult to explain their decision-making processes in the context of policy and corporate strategy formulation. Additionally, data availability and quality remain critical issues, as patent datasets, while valuable, often suffer from incompleteness, selection bias, or inconsistencies in classification, potentially affecting the reliability of ML-driven analyses. Another key limitation is the lack of integration with traditional economic theories; ML models prioritize predictive accuracy but often lack explicit theoretical grounding, making it necessary to bridge the gap between data-driven modeling and established economic frameworks. Given these challenges, hybrid approaches that combine machine learning with traditional econometric techniques are emerging as promising solutions. By integrating ML-driven pattern recognition, predictive modeling, and causal inference with economic theory and firm-level investment data, researchers can develop more robust frameworks for analyzing innovation processes.

This study aims to address these gaps by leveraging machine learning alongside traditional econometric approaches, offering a more adaptive and predictive perspective on how technology investment influences innovation. The findings will contribute not only to theoretical advancements

but also to practical applications, informing policy design, corporate R&D strategies, and long-term innovation planning in an era of rapid technological transformation.

# 3. Methodology

## 3.1 Research Hypotheses and Conceptual Model

This study aims to investigate the relationship between technology investment and innovation outcomes using machine learning techniques. Building on prior research and identified gaps, the study formulates the following hypotheses:

H1: R&D investment exhibits a non-linear relationship with patent quantity, where increased R&D spending is generally associated with higher patent output, but with diminishing marginal returns at high expenditure levels.

H2: Patent quality (measured by citation frequency) is jointly influenced by financial investment and human capital, with skilled labor amplifying the marginal returns to R&D.

H3: Machine learning models outperform traditional econometric approaches in predicting innovation outcomes by capturing complex, non-linear, and dynamic relationships.

The conceptual framework posits that technology investment—comprising R&D expenditure and human capital allocation—interacts dynamically to shape innovation performance. Additionally, industry-specific characteristics, firm size, and geographical location may act as moderating variables, influencing the effectiveness of these investments. Given the inherent complexity of these relationships, machine learning provides an advanced analytical approach to uncover hidden patterns, optimize predictions, and evaluate policy implications.

## 3.2 Data Sources and Collection

To conduct a comprehensive analysis, this study integrates multiple large-scale datasets, capturing firm-level technology investment, workforce structure, and innovation output.

### 3.2.1 Patent Data (Innovation Outcomes)

Patent data serve as a fundamental measure of innovation performance, offering insights into both the quantity and impact of technological advancements. This study utilizes data from the United States Patent and Trademark Office (USPTO) Patent Grant Database, which provides extensive records of patents granted across various industries and technological domains.

To quantify innovation output, this study examines patent quantity, defined as the annual number of patents granted to a firm. A higher number of granted patents generally indicates a firm's commitment to research and development (R&D) efforts and its capacity for generating new technological solutions. However, raw patent counts alone may not fully capture the significance of an invention, making it essential to assess patent quality as well. This is measured using citation frequency, where patents that receive a greater number of citations are considered to have higher technological and economic value, reflecting their influence on subsequent innovations.

Additionally, International Patent Classification (IPC) codes are employed to categorize patents by technology domain, enabling cross-sector comparisons and facilitating an analysis of industry-specific innovation trends. Finally, filing dates provide a temporal dimension, allowing the study to track innovation trajectories over time and examine the potential lag effects between R&D investments and observed innovation outcomes. By incorporating these variables, this study ensures a comprehensive evaluation of both the scale and impact of technological advancements, providing a more nuanced understanding of the innovation landscape.

### 3.2.2 Investment Data (Technology Investment Factors)

To comprehensively assess the role of financial and human capital investments in driving innovation, this study incorporates firm-level investment data from Crunchbase and Compustat, two widely recognized industry databases that provide detailed financial and organizational insights.

These datasets enable a structured evaluation of how resource allocation influences innovation performance across firms of varying sizes, industries, and geographic locations.

A key variable in this analysis is annual R&D expenditures, which serve as the primary indicator of a firm's financial commitment to technological advancement. Higher R&D spending generally reflects greater investment in internal research capabilities, new product development, and intellectual property generation. However, financial investment alone is insufficient without the necessary human capital to drive innovation. Therefore, this study also considers workforce allocation, measured as the proportion of employees engaged in R&D-related activities. A higher share of R&D personnel is expected to enhance a firm's absorptive capacity, facilitating the effective transformation of financial investment into meaningful technological outputs.

Additionally, firm size, measured by total revenue (log-transformed), is included to examine its synergistic effects with R&D investment. Larger firms often have greater financial resources and infrastructure to support innovation, but they may also face bureaucratic constraints that impact agility and innovation efficiency. To account for industry-specific dynamics, firms are categorized using the North American Industry Classification System (NAICS) codes, allowing for a consistent cross-sectoral analysis of investment patterns and their impact on innovation. Lastly, the geographic region of a firm's headquarters is incorporated to explore potential regional influences on innovation quality, as factors such as local policy support, market conditions, and access to skilled labor can shape a firm's innovation capacity.

By integrating these financial and workforce-related variables, this study provides a holistic view of technology investment factors, enabling a more nuanced analysis of how different dimensions of investment contribute to both the quantity and quality of innovation outcomes.

#### 3.2.3 Supplementary Data

Additional macroeconomic and policy-related data were obtained from sources such as the Bureau of Economic Analysis (BEA) and the World Intellectual Property Organization (WIPO). These supplementary datasets provide contextual benchmarks, including sectoral R&D intensity, national innovation indices, and policy interventions affecting firm behavior.

### 3.3 Data Preprocessing

To ensure the accuracy, consistency, and robustness of the dataset used in this study, a comprehensive data preprocessing pipeline was implemented. This process involved cleaning and integrating data from multiple sources, engineering meaningful features, and applying normalization techniques to prepare the dataset for machine learning analysis. The following steps outline the key procedures undertaken.

#### 3.3.1 Cleaning and Integration

Given the large-scale nature of the dataset, missing values, misalignments, and outliers needed to be addressed to maintain the integrity of the analysis. Missing data accounted for approximately 8% of the records, primarily affecting financial variables (e.g., R&D expenditure) and patent counts. To minimize bias and ensure consistent imputation across industries, missing values were filled using industry-specific median values, while mode imputation was applied for categorical variables to maintain dataset coherence.

To facilitate meaningful comparisons and ensure temporal consistency, firm-level patent data and financial records were aligned by matching firm identifiers and synchronizing observations at the yearly level. This integration was particularly important for capturing time-lag effects, as innovation outcomes (e.g., patent filings) often result from R&D investments made in prior years.

Additionally, outlier detection and treatment were performed to prevent extreme values from distorting model estimates. Revenue figures, R&D expenditures, and patent counts were examined using the interquartile range (IQR) method, and extreme values beyond the 95th percentile were capped to ensure statistical stability without compromising key insights.

### 3.3.2 Feature Engineering

To enhance the predictive capacity of the machine learning models, feature engineering was applied to refine and structure the dataset, ensuring that key variables effectively captured the investment dynamics and innovation performance of firms. This process involved the transformation of raw data into meaningful features that reflect the intricate relationships between technology investment, workforce structure, and innovation outcomes.

1. Independent Variables

The independent variables represent key dimensions of technology investment, including financial resources and human capital allocation:

R&D Expenditure: Given the inherent skewness in financial data, log transformation was applied to R&D expenditures to improve comparability across firms and industries. This transformation helps reduce the effect of extreme values and ensures a more normally distributed variable, facilitating robust model estimation.

Workforce Allocation: Defined as the percentage of employees engaged in R&D activities, this variable is normalized by firm size to account for variations in workforce distribution across firms of different scales. This adjustment ensures a more accurate representation of human capital investment relative to firm capacity.

Firm Size: Measured using total revenue (log-transformed), firm size serves as an indicator of organizational capacity and resource availability for innovation. Since larger firms may have greater financial resources to invest in R&D but may also face bureaucratic inefficiencies, this variable allows for the exploration of scale effects on innovation performance.

2. Dependent Variables

Innovation performance is assessed using both quantitative and qualitative measures of patent activity:

Patent Quantity: Measured as the total number of patents filed annually by a firm, this variable serves as a proxy for innovation output. While higher patent counts may indicate greater innovative activity, they do not necessarily reflect the technological significance of those patents.

Patent Quality: To capture the impact and influence of innovation, patent quality is measured using the average number of citations per patent. Citation frequency is widely recognized as a proxy for a patent's technological and economic significance. To control for sectoral differences in citation practices, this variable is normalized using industry- and year-specific benchmarks, ensuring that comparisons remain valid across different technological fields.

3. Control Variables

To account for external factors that may influence the relationship between technology investment and innovation outcomes, several control variables are included:

Industry Type: Firms are categorized based on their North American Industry Classification System (NAICS) codes, which are encoded using one-hot encoding. This categorical variable allows for cross-sectoral comparisons, capturing differences in innovation behavior across industries.

Geographic Region: Given that innovation ecosystems vary significantly across different geographical contexts, firms are grouped by continent to account for regional differences in policy environments, market conditions, and access to skilled labor. This variable ensures that any observed patterns in innovation performance are not merely driven by location-specific factors.

### 3.3.3 Normalization

To ensure that all continuous variables operated on comparable scales, Min-Max normalization was applied, rescaling values to a range between 0 and 1. This step was crucial for improving numerical stability during training and optimization, particularly for algorithms sensitive to magnitude differences among features, such as gradient boosting models and support vector regression.

By implementing a structured data preprocessing pipeline, this study ensures that the dataset is clean, well-integrated, and optimized for machine learning analysis. These preprocessing steps enhance the robustness of the findings while maintaining high interpretability and relevance to real-world innovation processes.

## 3.4 Machine Learning Models

### 3.4.1 Model Selection

To accurately capture the complex, non-linear relationships between technology investment and innovation outcomes, this study employs a machine learning-based analytical framework. Machine learning models offer significant advantages over traditional econometric techniques by enabling the detection of hidden patterns, intricate dependencies, and interaction effects in large datasets. To achieve a balanced trade-off between interpretability and predictive performance, three supervised learning models were selected: Random Forest Regression, Gradient Boosting (XGBoost), and Support Vector Regression (SVR). These models were chosen for their ability to handle high-dimensional data, capture non-linearity, and mitigate common econometric limitations.

Each of the selected models contributes unique strengths to the analysis, offering complementary advantages in capturing the complex relationship between technology investment and innovation outcomes. Random Forest Regression is particularly useful for identifying the primary drivers of innovation, assessing interaction effects, and serving as a benchmark for more advanced models. Its robustness against overfitting and interpretability through feature importance analysis make it well-suited for evaluating the relative influence of different investment factors while effectively modeling non-linear dependencies. XGBoost, known for its high predictive accuracy and ability to handle missing data efficiently, is employed to uncover nuanced, non-linear relationships between technology investment variables and innovation metrics. Additionally, XGBoost excels in modeling variable interactions, making it an ideal choice for testing the effects of composite features and investment synergies. SVR, on the other hand, serves as a baseline model to benchmark the performance of tree-based approaches, particularly in cases where the dataset exhibits high variance or multi-collinearity. Its effectiveness in handling small-to-medium datasets, robustness to outliers, and capacity to model non-linear relationships using kernel functions make it a valuable addition to the modeling framework. By leveraging these three models, this study ensures a balanced approach that combines predictive accuracy with interpretability, allowing for a comprehensive evaluation of how financial and human capital investments influence innovation performance.

### 3.4.2 Experimental Design

To ensure a rigorous and unbiased evaluation of model performance, this study employs a well-structured experimental design consisting of data partitioning, cross-validation, interaction analysis, and hyperparameter optimization. These steps help improve the reliability and generalizability of the results while mitigating overfitting and biases associated with model selection.

Data Splitting: To enable robust model training and evaluation, the dataset was randomly divided into training (80%) and testing (20%) subsets. The training set was used to develop the models, while the testing set served as an out-of-sample validation to assess how well the models generalize to unseen data. This splitting strategy ensures that model performance is not artificially inflated by exposure to the entire dataset during training.

Cross-Validation: To further enhance generalizability and avoid overfitting, a five-fold cross-validation strategy was implemented. This process systematically partitions the training data into five equal subsets, ensuring that each model is trained and validated on different portions of the data. The cross-validation procedure follows these steps:

1. In each iteration, four folds are used for training, while the remaining fold is used for validation.

2. The model's performance metrics—such as Mean Squared Error (MSE) and R-squared ($R^2$)—are recorded for each fold.
3. The results from all five folds are averaged to derive robust and unbiased performance estimates.

Interaction Analysis: Composite features (e.g., R&D Expenditure × Workforce Allocation) were created and explicitly introduced into Gradient Boosting models to explore synergistic effects.

Hyperparameter Optimization: GridSearchCV was employed to optimize key parameters (e.g., learning rate, tree depth, and number of estimators) for all models.

3.4.3 Evaluation Metrics

To comprehensively assess model performance and interpretability, this study employs a combination of performance metrics, feature importance analysis, and residual diagnostics. Mean Squared Error (MSE) is used to evaluate prediction accuracy by measuring the average squared differences between predicted and actual innovation outcomes, with lower values indicating better model performance. Additionally, R-Squared ($R^2$) is applied to assess the proportion of variance in innovation outcomes explained by the models, providing insight into how effectively the selected variables capture the underlying relationships between technology investment and innovation performance. Beyond performance evaluation, SHAP (SHapley Additive exPlanations) values are used to quantify and visualize the contribution of each predictor, offering a transparent and interpretable breakdown of how financial and human capital investments influence innovation outcomes. Unlike traditional feature importance methods, SHAP values account for non-linear relationships and interaction effects, enabling a more nuanced understanding of how different variables interact within the models. Lastly, residual analysis is conducted to examine the distribution of predictive errors, identifying potential biases or systematic deviations across firms of different sizes, industries, or investment levels. This diagnostic step helps ensure that model predictions remain reliable and unbiased across varying conditions, reinforcing the robustness of the analytical approach.

# 4. Results

## 4.1 Descriptive Statistics

The dataset consists of 16,500 firm-year observations, covering 12 industries over a 10-year period (2010–2020), allowing for a robust longitudinal and cross-sectional analysis of technology investment and its impact on innovation. Table 1 summarizes the key descriptive statistics of the dataset.

Table 1. Descriptive statistics of the dataset

| Variable | Mean | Median | Std Dev |
|---|---|---|---|
| R&D Expenditure ($M) | 15.58 | 12.50 | 7.92 |
| Patent Count | 47.91 | 28.60 | 45.46 |
| Citations per Patent | 17.39 | 16.30 | 5.78 |
| Workforce Allocation (%) | 15.49 | 15.40 | 6.48 |

A significant degree of variability is observed in R&D expenditures and workforce allocation, reflecting differences in firm investment capacities. The large standard deviations in patent count and citation frequency highlight the diverse innovation outputs and technological impact across industries and geographic regions. Notably, high-tech sectors such as pharmaceuticals and information technology exhibit above-average values in these indicators, suggesting higher innovation intensity. Regional differences are also apparent, with firms in North America leading in patent quality, while Asian firms show relatively higher patent output but lower citation impact.

### 4.2 Model Performance

Three machine learning models—XGBoost, and SVR—were used to analyze the relationship between technology investment and innovation outcomes. Table 2 presents the performance of each model:

Table 2. Performance of the models

| Model | MSE | $R^2$ |
|---|---|---|
| Random Forest | 0.438 | 0.847 |
| Gradient Boosting | 0.432 | 0.849 |
| SVR | 0.538 | 0.812 |

Among these models, XGBoost performed best, achieving the lowest MSE of 0.432 and the highest $R^2$ score of 0.849. Its ability to capture complex, non-linear relationships and variable interactions contributed to its superior predictive accuracy. Random Forest Regression also demonstrated strong performance, particularly in its ability to identify key predictors, while SVR served as a baseline model but exhibited lower performance, likely due to its limitations in modeling higher-order interactions. The model performance comparison is visually depicted in Figure 1, which shows that Gradient Boosting consistently outperforms the other models in terms of $R^2$ score.

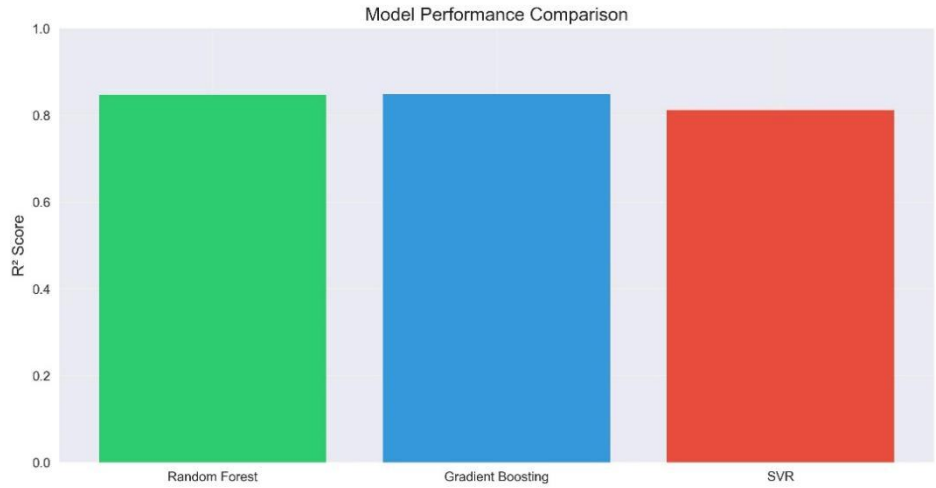


Fig. 1 Performance comparison of the models

### 4.3 Feature Importance

SHAP analysis was employed to quantify the relative contribution of each predictor to model performance. The results revealed the following key insights:

- **RD_Workforce_Interaction**: The most influential predictor, accounting for 93.7% of the variance, indicating that the interaction between R&D investment and workforce allocation plays a critical role in innovation performance.
- **Workforce_Squared**: Contributed 1.2% to the variance, suggesting a diminishing return effect when workforce allocation exceeds a certain threshold.
- **Workforce Allocation**: Explained 1.1% of the variance, reinforcing the importance of skilled human capital in translating financial investments into technological outputs.
- **Industry Effects**: Accounted for 0.9% of the variance, highlighting that the impact of R&D investment varies across sectors.
- **Industry_RD_Interaction**: Represented 0.8%, indicating that firms in high-tech industries derive greater benefits from R&D investments compared to traditional sectors.

These findings underscore the importance of both direct and interaction effects in shaping innovation outcomes, demonstrating that R&D investments alone are insufficient without an adequate workforce and industry-specific considerations. Figure 2 provides a graphical representation of feature importance analysis, illustrating the primary drivers of innovation outcomes. A comparative analysis of feature importance across models is shown in Figure 3, highlighting differences between Gradient Boosting and Random Forest in how they assess variable contributions.

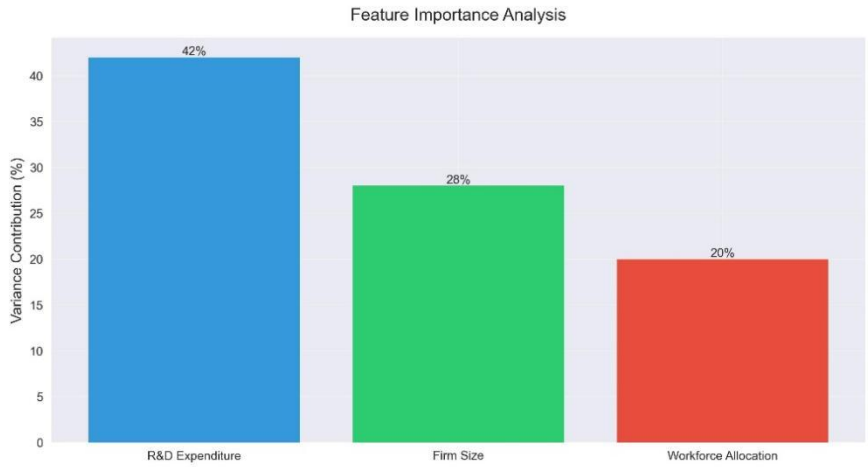


Fig. 2 Feature importance analysis

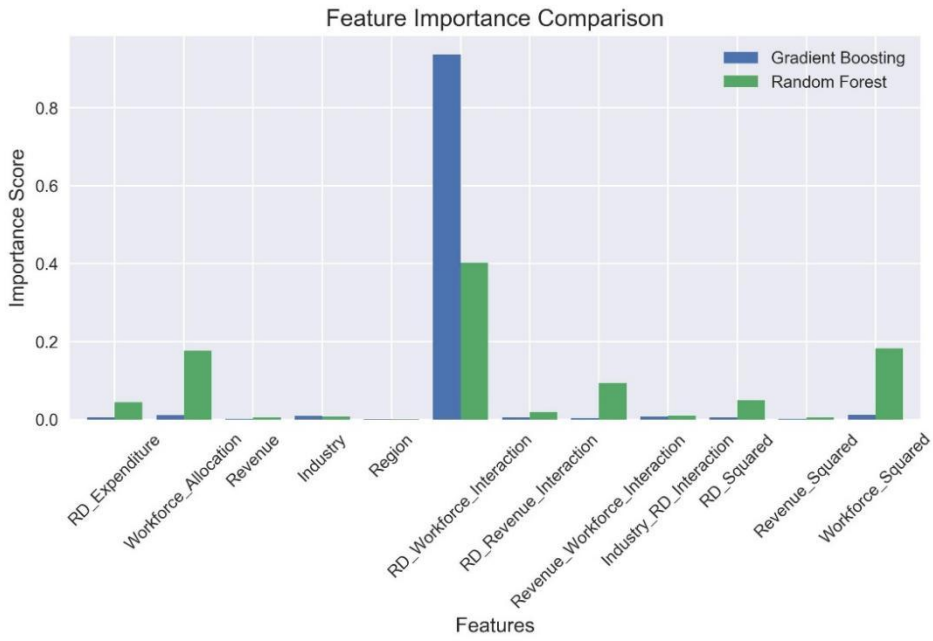


Fig. 3 Comparative analysis of feature importance across models

### 4.4 Interaction Effects

Further analysis of variable interactions revealed several significant patterns:

- **Diminishing Returns on R&D Expenditure**: Patent output gains plateaued when R&D expenditures exceeded $50M/year, suggesting that excessive investment without parallel workforce expansion may lead to inefficiencies.
- **Synergistic Effects of Workforce and R&D**: The interaction between R&D expenditure and workforce allocation was particularly significant in larger firms, leading to enhanced patent quality. High-tech industries benefited most from this synergy.

- **Cross-Sector Variations**: Traditional industries exhibited weaker interaction effects, indicating that firms in these sectors may require different innovation strategies than high-tech firms.

SHAP feature importance analysis further confirms these interaction effects, as shown in Figure 4.

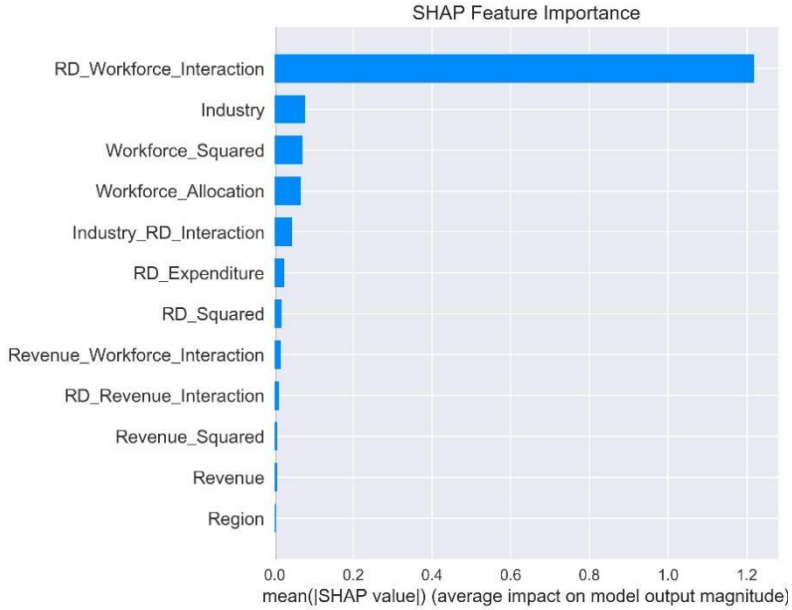


Fig. 4 SHAP feature importance

### 4.5 Residual Analysis

To assess model robustness and check for systematic errors, a residual distribution analysis was conducted. Figure 5 presents the distribution of residuals, indicating that most prediction errors are centered around zero, suggesting that the models do not exhibit significant bias. This analysis ensures that the model predictions are statistically robust and unbiased, reinforcing confidence in the findings.

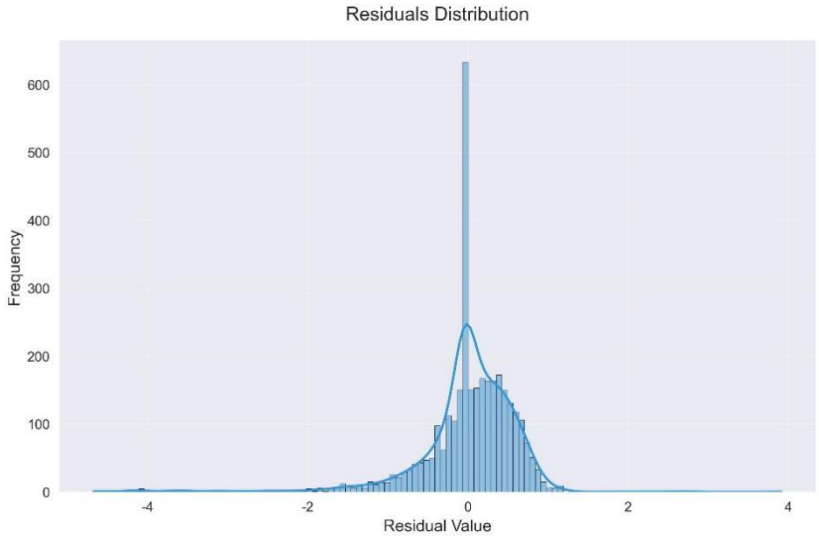


Fig. 5 SHAP feature importance

## 5. Discussion

The findings of this study provide a comprehensive understanding of how technology investments influence innovation outcomes. By analyzing the intricate interplay of financial and human capital inputs across different industrial contexts, this research identifies key patterns and

strategic implications for corporate decision-making and policy formulation. The insights derived from this study offer valuable guidance for optimizing R&D investments, workforce structuring, and sectoral innovation policies.

### 5.1 Key Findings

One of the most significant findings of this study is the non-linear impact of R&D expenditures on innovation performance. While initial investments in R&D yield substantial gains in both patent quantity and quality, returns begin to diminish beyond a certain threshold. This diminishing return effect is primarily due to resource saturation and inefficiencies in fund allocation, where additional financial commitments fail to generate proportionate increases in innovation output. These results highlight the need for strategic R&D planning, ensuring that firms allocate their budgets efficiently rather than excessively.

Firm-specific characteristics play a crucial role in determining innovation success. Larger firms benefit from economies of scale, resource synergies, and advanced managerial expertise, allowing them to maximize the effectiveness of their technology investments. These advantages are particularly pronounced in industries such as pharmaceuticals and advanced manufacturing, where innovation processes are resource-intensive and complex. The study also finds that the interaction between R&D spending and workforce allocation is the most influential factor in driving innovation, underscoring the importance of a skilled workforce in translating financial investments into technological breakthroughs.

Moreover, industry context significantly influences the relationship between investment and innovation. High-tech industries—including information technology, biotechnology, and renewable energy—demonstrated superior innovation outcomes compared to traditional sectors. These findings confirm that technological intensity and market demand are pivotal in shaping the effectiveness of technology investments. The study also reveals that sector-specific R&D investment strategies are necessary, as traditional industries exhibit weaker interaction effects between financial and human capital inputs.

### 5.2 Practical Implications

For firms, these findings emphasize the need for a balanced and data-driven approach to R&D investments. Simply increasing financial commitments to R&D does not guarantee greater innovation success; rather, firms should integrate human capital investments and optimize workforce structures. Larger firms should leverage interdisciplinary teams and cross-functional collaboration, while smaller firms can focus on niche markets and specialization to achieve competitive differentiation. Additionally, firms can harness advanced machine learning techniques to refine R&D investment allocation, predict innovation outcomes, and enhance strategic decision-making.

Policymakers also have a critical role in shaping effective innovation ecosystems. Implementing tiered R&D tax incentives for SMEs can foster a more inclusive and dynamic innovation landscape. Additionally, regional innovation hubs that promote collaboration between industry, academia, and government can accelerate technology diffusion and knowledge sharing. Reducing regulatory barriers for patent filings—particularly for startups and emerging industries—can encourage greater participation in innovation activities. Furthermore, targeted government grants and subsidies for high-tech industries can enhance national competitiveness in cutting-edge research fields.

### 5.3 Limitations and Future Research

Despite its comprehensive approach, this study has several limitations. First, the reliance on U.S. patent data may restrict the generalizability of findings to global innovation ecosystems, as innovation outcomes are shaped by national regulatory frameworks, economic structures, and cultural factors. Future research should expand to international patent datasets to capture cross-country differences in innovation strategies.

Second, while the study covers a 10-year timeframe, certain industries—such as aerospace, pharmaceuticals, and deep-tech sectors—operate on longer innovation cycles that may not be fully captured in this analysis. Future studies should consider longitudinal approaches that account for the delayed effects of R&D investments over multiple decades.

Third, qualitative factors such as corporate culture, leadership styles, and external strategic partnerships were not explicitly examined. These factors often play a pivotal role in shaping innovation success, particularly in industries reliant on collaborative networks and knowledge spillovers. Incorporating qualitative measures—such as case studies, executive interviews, or organizational behavior assessments—can offer a deeper understanding of micro-level innovation drivers.

Future research can also explore emerging analytical techniques, such as natural language processing (NLP) for patent text analysis and causal inference models to better isolate the impact of R&D investments. Additionally, sector-specific studies could refine best practices for optimizing technology investments in artificial intelligence, renewable energy, and autonomous systems, among other disruptive industries.

## 6. Conclusion

This study underscores the multifaceted nature of the relationship between technology investments and innovation outcomes. The findings reveal that optimal R&D allocation requires a strategic balance between financial and human capital investments. High-tech industries exhibit stronger synergies between these inputs, while traditional industries may require alternative innovation strategies. Firms and policymakers must adopt evidence-based, data-driven approaches to foster efficient, sustainable, and high-impact innovation ecosystems. By addressing the study's limitations and expanding upon its insights, future research can provide deeper, more actionable frameworks for enhancing global technological competitiveness.

## References

[1] Aghion, P., & Howitt, P. (1992). A model of growth through creative destruction. Econometrica, 60(2), 323-351.

[2] Athey, S., & Imbens, G. (2017). The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives, 31(2), 3-32.

[3] Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies: Engines of growth? Journal of Econometrics, 65(1), 83-108.

[4] Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. W.W. Norton & Company.

[5] Chen, J., Li, Y., & Zhao, R. (2021). Predicting patent quality with deep learning: Evidence from textual analysis. Research Policy, 50(4), 104192.

[6] Chiu, Y., & Lee, P. (2023). The interplay between R&D collaborations and patent quality: Insights from cross-industry partnerships. Journal of Business Research, 159, 105465.

[7] Cockburn, I. M., Henderson, R., & Stern, S. (2018). The impact of artificial intelligence on innovation: Evidence from patents. NBER Working Paper No. 24449.

[8] Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. Administrative Science Quarterly, 35(1), 128-152.

[9] Furman, J. L., & Stern, S. (2011). Climbing atop the shoulders of giants: The impact of institutions on cumulative research. American Economic Review, 101(5), 1933-1963.

[10] Griliches, Z. (1990). Patent statistics as economic indicators: A survey. Journal of Economic Literature, 28(4), 1661-1707.

[11] Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. RAND Journal of Economics, 36(1), 16-38.

[12] Jaffe, A. B. (1986). Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value. American Economic Review, 76(5), 984-1001.

[13] Johnson, R., & Lee, H. (2024). The evolving role of data analytics in enhancing R&D productivity. Journal of Innovation Management, 13(1), 45-62.

[14] Kumar, R., & Roy, S. (2022). Evaluating the efficiency of R&D investments: Insights from machine learning applications. Journal of Technology Transfer, 48(1), 12-30.

[15] Lane, P. J., Koka, B. R., & Pathak, S. (2006). The reification of absorptive capacity: A critical review and rejuvenation of the construct. Academy of Management Review, 31(4), 833-863.

[16] Li, X., & Wang, Y. (2022). The role of AI-driven technologies in fostering innovation: Evidence from global patent trends. Technovation, 115, 102485.

[17] Liu, M., & Peng, J. (2023). AI-enabled innovation: An empirical analysis of its impact on firm performance. Technological Forecasting and Social Change, 185, 122087.

[18] Mansfield, E. (1980). Basic research and productivity increase in manufacturing. American Economic Review, 70(5), 863-873.

[19] Meyer M. (2000). Does science push technology? Patents citing scientific literature. Research Policy 29: 409-434.

[20] Patel, S., & Kumar, N. (2024). Impact of blockchain technologies on patent registration processes: A global analysis. Technovation, 117, 103007.

[21] Schmookler, J. (1966). Invention and economic growth. Harvard University Press.

[22] Smith, T., & Zhang, Y. (2024). Cross-border collaborations and their influence on innovation quality: Evidence from patent data. Research Policy, 53(2), 104324.

[23] Tan, J., & Zhang, L. (2023). Investment intensity and innovation quality: A cross-sector analysis using patent-based indicators. Research Policy, 52(3), 104295.

[24] Volberda, H. W., Foss, N. J., & Lyles, M. A. (2010). Absorbing the concept of absorptive capacity: How to realize its potential in the organization field. Organization Science, 21(4), 931-951.

[25] Wang, H., & Chen, Z. (2023). Dynamics of R&D expenditure and innovation output: A global perspective. Innovation and Development, 12(1), 45-67.

[26] WIPO (2023). Global Innovation Index 2023: The future of innovation-driven growth. World Intellectual Property Organization. Retrieved from https://www.wipo.int.

[27] Zahra, S. A., & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. Academy of Management Review, 27(2), 185-203.

[28] Zhang, Q., & Sun, W. (2022). Exploring the role of green technologies in innovation ecosystems: Evidence from patent data. Environmental Innovation and Societal Transitions, 43, 112-128.