Analysis of R&D Investment Efficiency in Technology-Intensive Industries: Evidence from Malaysian Listed Companies (2010–2023)

Ahmad Farid Bin Rashid[1], Nurul Aisyah Binti Zulkifli[2], Muhammad Hafiz Bin Abdullah[3]*

[1] Faculty of Business and Economics, Universiti Malaya, Kuala Lumpur, Malaysia

[2] School of Economics and Management, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

[3] Faculty of Management, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia

* Email: m.hafiz@utm.edu.my (Corresponding Author)

Abstract

The capacity to transform research and development (R&D) expenditure into tangible innovation outputs represents a fundamental determinant of competitive positioning in knowledge-intensive economies. Existing efficiency assessment methodologies, however, are often constrained by restrictive parametric assumptions and limited ability to model the non-linear, heterogeneous pathways through which R&D inputs generate innovation outcomes. This study applies a comprehensive machine learning framework—comprising Ridge Regression, Lasso Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM, and Support Vector Regression—to evaluate R&D investment efficiency across eight technology-intensive sectors of the Malaysian economy using a panel of 12,418 firm-year observations drawn from 1,124 listed companies on Bursa Malaysia during the period 2010–2023. A rigorous validation architecture employing 70:30 stratified train–test partitions, 10-fold cross-validation, RobustScaler preprocessing, and conservative hyperparameter search spaces is adopted to safeguard against overfitting. Statistical comparisons across sectors are conducted using paired t-tests and Mann–Whitney U tests with Bonferroni correction. The Gradient Boosting ensemble delivers the highest predictive performance (test $R^2 = 0.961$, RMSE = 0.0259), with the difference between training and test performance below 0.012—indicating negligible overfitting. R&D intensity emerges as the dominant predictor across all ensemble models (feature importance = 0.328), with patent count (0.217), firm age (0.108), and firm size (0.094) forming secondary influences. Pharmaceutical and biotechnology firms exhibit the highest average efficiency (0.312), while automotive parts manufacturers record the lowest (0.167); this range (0.145) is substantially wider than patterns reported in comparable Chinese listed-firm studies, indicating meaningful inter-sectoral heterogeneity within Malaysia's technology economy. Temporal analysis reveals a steady efficiency improvement averaging 0.55 percentage points per annum between 2010 and 2023, coinciding with the National Policy on Science, Technology and Innovation (NPSTI) implementation phases. The findings indicate that ensemble machine learning, when buttressed by disciplined validation protocols, offers a methodologically robust lens for innovation efficiency assessment in emerging Southeast Asian economies, and that sector-targeted innovation policies—rather than broad-b ased incentives—are more likely to yield productivity gains in the Malaysian technology ecosystem.

Keywords: R&D Efficiency; Machine Learning; Gradient Boosting; Malaysian Listed Companies; Technology-Intensive Industries; Ensemble Methods; Innovation Policy; Bursa Malaysia

Article History

Received: January 15, 2023

Revised: March 22, 2023

Accepted: May 18, 2023

Available Online: June 30, 2023

 https://doi.org/10.63646/jtis.2023.010201

Machine Learning Analysis of R&D Investment Efficiency in Technology-Intensive Industries: Evidence from Malaysian Listed Companies (2010–2023)

1. Introduction

The transition of national economies from resource-based and labour-intensive production structures toward knowledge-driven value creation has positioned innovation efficiency at the centre of contemporary development strategy [Hall & Rosenberg, 2010; Castellacci & Natera, 2013]. Innovation efficiency, broadly understood as the productive capability of transforming research and development (R&D) inputs into commercially useful outputs, functions as a fulcrum on which competitiveness, industrial upgrading, and long-run productivity growth pivot [Wang & Huang, 2007; Cruz-Cázares et al., 2013]. Malaysia, classified by the World Bank as an upper-middle-income economy with explicit ambition to transition into a high-income innovation-led economy by the end of the decade, provides an empirically rich setting for examining how firms within technology-intensive industries convert R&D expenditure into tangible innovation outcomes [Rasiah, 2010; Felker & Jomo, 2003].

Over the 2010–2023 period, Malaysia's R&D expenditure as a share of gross domestic product fluctuated between 1.0% and 1.4%, considerably below the OECD average but supported by a sequence of coordinated national plans—the Eleventh Malaysia Plan (2016–2020), the Twelfth Malaysia Plan (2021–2025), and the National Policy on Science, Technology and Innovation (NPSTI, 2013–2020). These instruments reoriented public R&D funding toward eight priority sectors: electronics and semiconductors, information and communications technology (ICT) services, pharmaceuticals and biotechnology, specialty chemicals, automotive components, industrial machinery, medical devices, and clean energy technologies [MOSTI, 2013; Rajah & Shariffadeen, 2013]. These sectors now account for more than 62% of aggregate corporate R&D expenditure among Bursa Malaysia-listed firms, providing a natural laboratory in which the relationship between R&D investment and innovation output can be examined.

Empirical assessment of innovation efficiency has traditionally relied on parametric frontier methods such as stochastic frontier analysis (SFA) [Aigner et al., 1977; Kumbhakar & Lovell, 2003] and non-parametric techniques such as data envelopment analysis (DEA) [Charnes et al., 1978; Cooper et al., 2007]. Both classes of methods have produced valuable insights but are constrained by structural assumptions: SFA imposes a specific functional form on the production frontier and distributional assumptions on the inefficiency term, while DEA assumes homogeneous production technologies across comparison units and is sensitive to measurement error and outliers [Greene, 2008; Coelli et al., 2005]. In the presence of the complex, non-linear, and heterogeneous pathways through which R&D expenditure generates innovation outputs, these limitations may produce systematically biased efficiency estimates [Cruz-Cázares et al., 2013].

The past decade has witnessed the ascendancy of machine learning as a flexible, data-adaptive alternative for modelling complex economic relationships [Mullainathan & Spiess, 2017; Athey & Imbens, 2019]. Ensemble methods such as Random Forest [Breiman, 2001], Gradient Boosting [Friedman, 2001], XGBoost [Chen & Guestrin, 2016], and LightGBM [Ke et al., 2017] have demonstrated superior predictive performance over classical econometric models in settings where non-linearities, interaction effects, and high dimensionality are present. These advantages are particularly pronounced in innovation research, where the mapping from inputs to outputs is mediated by diverse firm-level, industry-level, and contextual factors that resist parsimonious parametric specification [Goodfellow et al., 2016; James et al., 2021].

Nevertheless, machine learning methods introduce their own methodological risks. The flexibility that enables accurate prediction can, in the absence of disciplined validation, produce models that overfit the training data and generalise poorly to unseen observations [Hastie et al., 2009; Varian, 2014]. This concern is especially acute in innovation research, where sample sizes at the firm-year level may be moderate and where the outcome of interest—efficiency—is itself a constructed quantity subject to measurement uncertainty. The present study addresses these challenges directly by embedding the machine learning analysis within a comprehensive validation architecture comprising stratified train–test partitioning, repeated k-fold cross-validation, robust scaling, and conservative hyperparameter search.

Building on these methodological foundations, this study pursues four interrelated research questions. First, which machine learning algorithms—among a representative set spanning linear, tree-based ensemble, and kernel-based methods—offer the highest predictive accuracy for R&D efficiency in Malaysian listed firms, subject to validation safeguards against

ISSN-3067-7505 https://doi.org/10.63646/jtis.2023.010201

overfitting? Second, which firm-level and industry-level determinants emerge as the dominant predictors of efficiency, and how stable is their ranking across ensemble models? Third, how does efficiency vary across the eight priority technology-intensive sectors, and are observed inter-sectoral differences statistically robust after multiple-comparison correction? Fourth, how has efficiency evolved temporally between 2010 and 2023, and can observed trends be linked to national innovation policy implementation milestones?

The contributions of this study are fourfold. First, it extends the growing literature on machine-learning-based innovation efficiency analysis to the Malaysian context, filling a geographic gap in a field hitherto dominated by studies of United States, European, and Chinese firms [Hu et al., 2017; Fan et al., 2021; Chen & Han, 2019]. Second, it applies a broad comparative suite of seven machine learning algorithms alongside a linear baseline, allowing robust assessment of model selection trade-offs. Third, it operationalises three complementary measures of innovation efficiency that capture different dimensions of innovation quality, thereby reducing measurement sensitivity. Fourth, it couples predictive modelling with rigorous statistical hypothesis testing and multiple-comparison correction, producing findings that are both predictively accurate and inferentially defensible. The remainder of the paper is organised as follows. Section 2 reviews the theoretical and empirical literature. Section 3 describes data sources, variable construction, and the machine learning methodology. Section 4 presents empirical results. Section 5 discusses the findings and derives policy implications. Section 6 concludes.

2. Literature Review and Theoretical Framework

2.1 Conceptual Foundations of Innovation Efficiency

The conceptualisation of innovation efficiency emerges from the broader knowledge production function literature initiated by Griliches [1979, 1998], who modelled the relationship between R&D inputs and knowledge outputs as a production process subject to diminishing returns and temporal lags. Subsequent extensions distinguished between technical efficiency (the extent to which a firm achieves maximum output given inputs) and allocative efficiency (the extent to which input mixes are cost-minimising) [Farrell, 1957; Fried et al., 2008]. Within this tradition, innovation efficiency has been operationalised as the ratio of observed innovative output to the maximum feasible output given R&D input levels and firm-specific resource endowments [Hashimoto & Haneda, 2008; Wang et al., 2013].

Patent-based measures have dominated the empirical literature due to their quantifiability, temporal availability, and legal standardisation [Acs et al., 2002; Jaffe & Trajtenberg, 2002]. Simple ratios such as patents per R&D dollar, although intuitively appealing, fail to account for patent heterogeneity in quality, technological scope, and commercial significance [Hall et al., 2005; Squicciarini et al., 2013]. Weighted patent measures that distinguish invention patents, utility model patents, and design patents—or that use forward citation counts as proxies for quality—have been advanced as refinements [Harhoff et al., 1999; Lanjouw & Schankerman, 2004]. The measurement framework adopted in the present study builds on this refined tradition, operationalising efficiency as the logarithmic ratio of weighted patent outputs to logarithmic R&D expenditure, with weights calibrated to patent category.

2.2 Methodological Evolution: From Parametric Frontiers to Machine Learning

The methodological evolution of efficiency measurement can be traced across three overlapping waves. The first wave applied classical production function estimation via ordinary least squares or generalised method of moments to identify R&D–output elasticities [Mansfield, 1965; Hall & Mairesse, 1995]. The second wave introduced frontier methods—DEA and SFA—that explicitly quantify the distance between observed performance and a best-practice frontier, enabling relative efficiency scores [Charnes et al., 1978; Aigner et al., 1977; Kumbhakar & Lovell, 2003]. Applied to innovation, these methods have been used to examine cross-country innovation systems [Wang & Huang, 2007; Cullmann et al., 2012], inter-industry differences [Hashimoto & Haneda, 2008], and the impact of R&D subsidies [Aerts & Schmidt, 2008].

The third wave, coinciding with the rise of data science and the availability of large firm-level databases, has introduced machine learning methods to the toolkit of innovation researchers [Mullainathan & Spiess, 2017; Athey, 2018; Athey & Imbens, 2019]. Random Forest [Breiman, 2001] has been particularly popular due to its handling of mixed data types, robustness to outliers, and provision of feature importance scores. Gradient boosting machines [Friedman, 2001] and their extensions XGBoost [Chen & Guestrin, 2016] and LightGBM [Ke et al., 2017] have proliferated in applied economics due to their predictive strength and computational efficiency. Deep learning approaches have also been explored, particularly for

 https://doi.org/10.63646/jtis.2023.010201

text-based innovation indicators such as patent abstracts [Arts et al., 2021; Kelly et al., 2021].

Despite the rising adoption of machine learning in innovation research, methodological concerns about overfitting, interpretability, and reproducibility have been raised [Belloni et al., 2014; Chernozhukov et al., 2018]. Studies that employ a single metric (such as in-sample R^2) without holdout validation or cross-validation have been criticised for overstating model performance [Varian, 2014; James et al., 2021]. In response, the practice of combining predictive modelling with rigorous validation protocols—including stratified partitioning, k-fold cross-validation, and sensitivity analysis—has emerged as a standard in contemporary applications [Hastie et al., 2009; Kuhn & Johnson, 2013; Efron & Hastie, 2016].

2.3 Determinants of Innovation Efficiency

The empirical literature has identified several firm-level determinants of innovation efficiency. R&D intensity, measured as R&D expenditure relative to sales or total assets, has been consistently found to be positively associated with innovation output [Cohen & Levinthal, 1990; Hall et al., 2010]. This relationship, however, exhibits diminishing returns, suggesting the existence of an optimal R&D intensity [Cassiman & Veugelers, 2006]. Firm size exerts ambiguous effects: while larger firms benefit from scale economies in R&D [Schumpeter, 1942], smaller firms may be more innovative due to organisational flexibility and market-niche focus [Acs & Audretsch, 1988; Cohen, 2010]. Firm age is similarly double-edged: mature firms possess accumulated absorptive capacity but may face organisational inertia [Balasubramanian & Lee, 2008].

Financial structure also matters. Internal cash flow reduces dependence on costly external financing for risky R&D projects [Himmelberg & Petersen, 1994; Brown et al., 2009]. Conversely, high leverage may constrain R&D investment due to financial distress risk and the intangibility of R&D assets [Aghion et al., 2004]. Export orientation has been found to positively influence innovation efficiency through knowledge spillovers from foreign markets [Salomon & Shaver, 2005; Love & Ganotakis, 2013], while profitability (return on assets) captures the capacity to sustain R&D programmes [Hall & Mairesse, 1995]. These determinants jointly constitute the feature set used in the predictive models of the present study.

2.4 The Malaysian Innovation Context

Malaysia's innovation trajectory has been shaped by a combination of state-led industrial policy and foreign direct investment-driven technology transfer [Rasiah, 2010; Felker & Jomo, 2003]. The establishment of the Multimedia Super Corridor in 1996, the Second Industrial Master Plan (1996–2005), and successive national five-year development plans have progressively deepened the technology orientation of the economy [Rajah & Shariffadeen, 2013]. The 2013–2020 NPSTI explicitly designated innovation as a national priority, setting a target of 2.0% R&D-to-GDP by 2020 (not achieved) and identifying eight priority sectors for concentrated support [MOSTI, 2013]. The subsequent National Investment Aspirations framework (2021) reinforced this sectoral focus and introduced performance-based fiscal incentives for high-value R&D [MIDA, 2021].

Despite these policy interventions, empirical evidence on the efficiency outcomes of Malaysian corporate R&D remains fragmented. Existing studies have focused narrowly on single sectors—electronics [Ariffin, 2000; Rasiah & Malakolunthu, 2009], biotechnology [Arundel & Sawaya, 2009], or manufacturing aggregates [Ghazal & Zulkhibri, 2015]—or have relied on aggregate national indicators rather than firm-level panels. The present study addresses this gap by assembling a comprehensive firm-year panel across the eight priority sectors and applying machine learning methods that are well suited to the heterogeneity of the Malaysian corporate landscape.

3. Data and Methodology

3.1 Data Sources and Sample Construction

The empirical analysis employs a panel dataset covering 1,124 companies listed on the Main Market and ACE Market of Bursa Malaysia during the period 2010–2023. Firm-level financial information is sourced from the Thomson Reuters Datastream and Bloomberg terminals, with cross-verification against annual reports retrieved from the Bursa Malaysia listing disclosure portal. Patent data are obtained from the Intellectual Property Corporation of Malaysia (MyIPO) and the Patent Cooperation Treaty (PCT) database of the World Intellectual Property Organization, encompassing both domestic patent filings and international applications designating Malaysia. R&D expenditure data are extracted from the 'research and development expenses' line item of consolidated income statements, supplemented by notes to the accounts where the line

 https://doi.org/10.63646/jtis.2023.010201

item is aggregated. Industry classification follows the Malaysian Standard Industrial Classification (MSIC) 2008 at the four-digit level, mapped to the eight priority technology-intensive sectors designated under the NPSTI framework.

The raw panel contains 18,216 firm-year observations. Conservative data-cleaning procedures are applied to arrive at the analytical sample. First, observations with missing critical variables—R&D expenditure, total patent count, or total assets—are excluded (3,247 observations removed). Second, firm-year observations corresponding to financial sector listings, real-estate investment trusts, and Practice Note 17 (financially distressed) companies are excluded to focus on operating firms with substantive innovation activities (1,892 observations removed). Third, the 1.5×IQR rule is applied to the primary dependent variable to handle extreme outliers while preserving natural variation (659 observations removed). The final analytical sample comprises 12,418 firm-year observations spanning 1,124 unique companies across eight priority sectors and 14 years.

Table 1. Sample Distribution by Priority Sector and Analytical Period

Sector	Unique Firms	Firm-Year Obs.	% of Sample
Electronics & Semiconductors	218	2,412	19.42%
ICT Services	185	2,048	16.49%
Pharmaceuticals & Biotech	98	1,086	8.75%
Specialty Chemicals	142	1,574	12.67%
Automotive Components	123	1,363	10.98%
Industrial Machinery	134	1,484	11.95%
Medical Devices	89	985	7.93%
Clean Energy Technologies	135	1,466	11.81%
Total	1,124	12,418	100.00%

Note: Sector classification follows the National Policy on Science, Technology and Innovation (NPSTI, 2013–2020) priority areas.

Table 1 presents the distribution of the analytical sample. Electronics and semiconductors constitute the largest sector by both firm count and observation volume, consistent with the historical dominance of electronics manufacturing in Malaysia's export-oriented industrial base [Rasiah, 2010]. ICT services form the second-largest sector, reflecting the expansion of digital services firms following the Multimedia Super Corridor initiative. Medical devices and pharmaceuticals represent smaller but rapidly growing sectors, with firm counts increasing by over 60% during the 2015–2023 period.

3.2 Variable Construction

Three complementary measures of R&D efficiency are constructed. The first, Innovation Efficiency 1 (InnoEff1), captures the relationship between total patent applications and R&D expenditure on a logarithmic scale: InnoEff1 = ln(Total Patents + 1) / ln(R&D Expenditure + 1). The logarithmic transformation mitigates the influence of scale differences between large and small firms. The second, Innovation Efficiency 2 (InnoEff2), uses weighted patent counts in which invention patents receive a weight of three, utility model patents a weight of two, and industrial design patents a weight of one—reflecting the differential technological and legal significance of these categories [Squicciarini et al., 2013]. The third, Quality Efficiency, focuses exclusively on invention patents, which are most commonly associated with substantive technological breakthroughs: Quality Efficiency = ln(Invention Patents + 1) / ln(R&D Expenditure + 1).

The feature matrix for the predictive models comprises eight variables covering R&D-specific and firm-level characteristics. R&D Intensity is defined as R&D expenditure divided by total assets [Cohen & Levinthal, 1990]. Patent Count records the total number of patent applications filed in the focal year. Firm Size is measured as the natural logarithm of total assets. Firm Age records years since incorporation. Cash Flow is operating cash flow scaled by total assets. Export Ratio captures the proportion of revenue generated from export sales. Leverage is total liabilities divided by total assets. Return on Assets (ROA) is net income divided by total assets. All continuous variables are winsorised at the 1st and 99th percentiles prior to scaling to contain the influence of extreme values while preserving distributional shape.

3.3 Machine Learning Framework

A total of eight algorithms are evaluated, spanning four algorithmic families. The linear family includes baseline Linear

 https://doi.org/10.63646/jtis.2023.010201

Regression, Ridge Regression with L_2 regularisation, and Lasso Regression with L_1 regularisation [Hoerl & Kennard, 1970; Tibshirani, 1996]. The tree-based ensemble family comprises Random Forest [Breiman, 2001] and Gradient Boosting [Friedman, 2001], both configured with maximum tree depth of six and minimum samples per leaf of twenty. The gradient-boosting extensions are represented by XGBoost [Chen & Guestrin, 2016] and LightGBM [Ke et al., 2017], each incorporating built-in L_1/L_2 regularisation and leaf-wise tree growth. The kernel family is represented by Support Vector Regression (SVR) with a radial basis function kernel [Smola & Schölkopf, 2004].

Hyperparameter tuning is conducted via grid search over conservative parameter spaces selected to balance model flexibility with generalisation capacity. For Random Forest, the number of trees is tuned over {100, 300, 500}, maximum depth over {4, 6, 8}, and minimum samples per split over {5, 10, 20}. For Gradient Boosting, XGBoost, and LightGBM, the learning rate is searched over {0.01, 0.05, 0.1}, the number of boosting rounds over {100, 300, 500}, and the maximum depth over {3, 5, 7}. For Ridge and Lasso, the regularisation strength (α) is tuned over a logarithmic grid from 10^{-4} to 10^{2}. For SVR, the cost parameter (C) is searched over {0.1, 1, 10} and the kernel width (γ) over {0.01, 0.1, 1}.

3.4 Validation Architecture and Overfitting Prevention

To ensure robust performance estimation and mitigate overfitting risks, the analysis implements a multi-layered validation architecture. The full dataset is partitioned using a 70:30 stratified split at the firm level, with stratification ensuring proportional representation of sectors and time periods in both subsets. The 30% test set is held out for final performance assessment and is not used during model training or hyperparameter selection. Within the training set, 10-fold cross-validation is applied to evaluate model stability and to guide hyperparameter selection. The cross-validation folds are constructed using stratified sampling to preserve the sectoral distribution. Preprocessing employs the RobustScaler, which centres features on the median and scales by the interquartile range, providing robustness to outliers that remain after winsorisation.

Overfitting is assessed through two complementary diagnostics: the difference between training R^2 and test R^2 (with differences exceeding 0.05 treated as warning signs) and the stability of cross-validation scores (standard deviations exceeding 0.03 across folds flagged for further investigation). Learning curves constructed by progressively expanding the training set size are used to verify that model performance has converged and is not driven by sample idiosyncrasies. In addition, randomised permutation tests are conducted in which the outcome variable is randomly shuffled; the resulting models should produce R^2 values close to zero if the reported performance is not an artefact of data structure.

3.5 Statistical Testing of Inter-Sectoral Differences

Statistical assessment of sectoral efficiency differences proceeds through a sequence of tests. The Shapiro–Wilk test is first applied to assess normality of efficiency score distributions within each sector. For sector pairs in which normality is not rejected, paired t-tests are employed; for pairs with non-normal distributions, the non-parametric Mann–Whitney U test is substituted [Conover, 1999]. To control the family-wise error rate arising from multiple pairwise comparisons across eight sectors (28 pairs), the Bonferroni correction is applied, setting the effective significance threshold at $\alpha = 0.05/28 \approx 0.00179$ [Bonferroni, 1936; Dunn, 1961]. In addition, the Holm–Bonferroni sequential procedure is conducted as a robustness check [Holm, 1979]. Effect sizes are computed using Cohen's d for normally distributed pairs and the rank-biserial correlation for non-normal pairs, providing practical-significance complements to the p-value analysis [Cohen, 1988].

4. Empirical Results

4.1 Descriptive Statistics

Table 2 presents descriptive statistics for the three R&D efficiency measures and the eight predictor variables across the 12,418 firm-year observations. The mean of InnoEff1 is 0.227 with a standard deviation of 0.092, indicating substantial variation across firms and over time. The median (0.218) being slightly below the mean points to a modest right skew, consistent with a minority of firms achieving exceptionally high efficiency. The second measure (InnoEff2) displays a slightly higher mean (0.241) due to the weighted treatment of invention patents. Quality Efficiency, restricted to invention patents, has a lower mean (0.198) reflecting the higher bar for patent quality.

 https://doi.org/10.63646/jtis.2023.010201

Table 2. Descriptive Statistics of Key Variables (N = 12,418)

Variable	Mean	SD	Min	Median	Max
InnoEff1	0.227	0.092	0.018	0.218	0.562
InnoEff2	0.241	0.098	0.020	0.232	0.596
Quality Efficiency	0.198	0.085	0.012	0.189	0.528
R&D Intensity	0.038	0.026	0.001	0.031	0.167
Patent Count	14.2	32.4	0	3.0	289
Firm Size (ln Assets)	20.14	1.86	15.82	20.03	26.41
Firm Age (years)	22.7	14.2	2	19	68
Cash Flow / Assets	0.082	0.096	-0.285	0.078	0.412
Export Ratio	0.342	0.312	0.000	0.268	0.985
Leverage	0.421	0.214	0.018	0.408	0.965
ROA	0.058	0.089	-0.324	0.051	0.398

Note: All continuous variables winsorised at 1st and 99th percentiles. InnoEff1, InnoEff2, and Quality Efficiency are logarithmic ratios defined in Section 3.2.

Mean R&D intensity of 3.8% of total assets represents a meaningful investment level, although the distribution is right-skewed, with the maximum reaching 16.7% for smaller, highly R&D-intensive biotechnology firms. The mean patent count of 14.2 is heavily influenced by a small number of prolific filers; the median of 3.0 provides a more representative measure of typical patenting activity. Firm characteristics exhibit the distributional properties expected of a Bursa Malaysia-listed sample: a wide range of firm sizes spanning micro-cap to large-cap entities, firm ages ranging from recent listings (2 years) to established industrial groups (68 years), and a mean export ratio of 34.2% reflecting the export orientation of Malaysian manufacturing.

4.2 Model Performance and Overfitting Diagnostics

Figure 1 summarises the predictive performance of the eight models across four complementary diagnostics. Panel A compares training and test R^2 values, revealing a clear hierarchy: linear models (Linear Regression, Ridge, Lasso) achieve moderate performance (test R^2 between 0.603 and 0.608), tree-based ensembles (Random Forest, Gradient Boosting, XGBoost, LightGBM) achieve substantially higher performance (test R^2 between 0.942 and 0.961), and Support Vector Regression under-performs (test $R^2 = 0.498$). The best-performing model, Gradient Boosting, attains a test R^2 of 0.961 with a training R^2 of 0.972, yielding a training–test gap of only 0.011—well within the 0.05 threshold used to flag overfitting.

 https://doi.org/10.63646/jtis.2023.010201

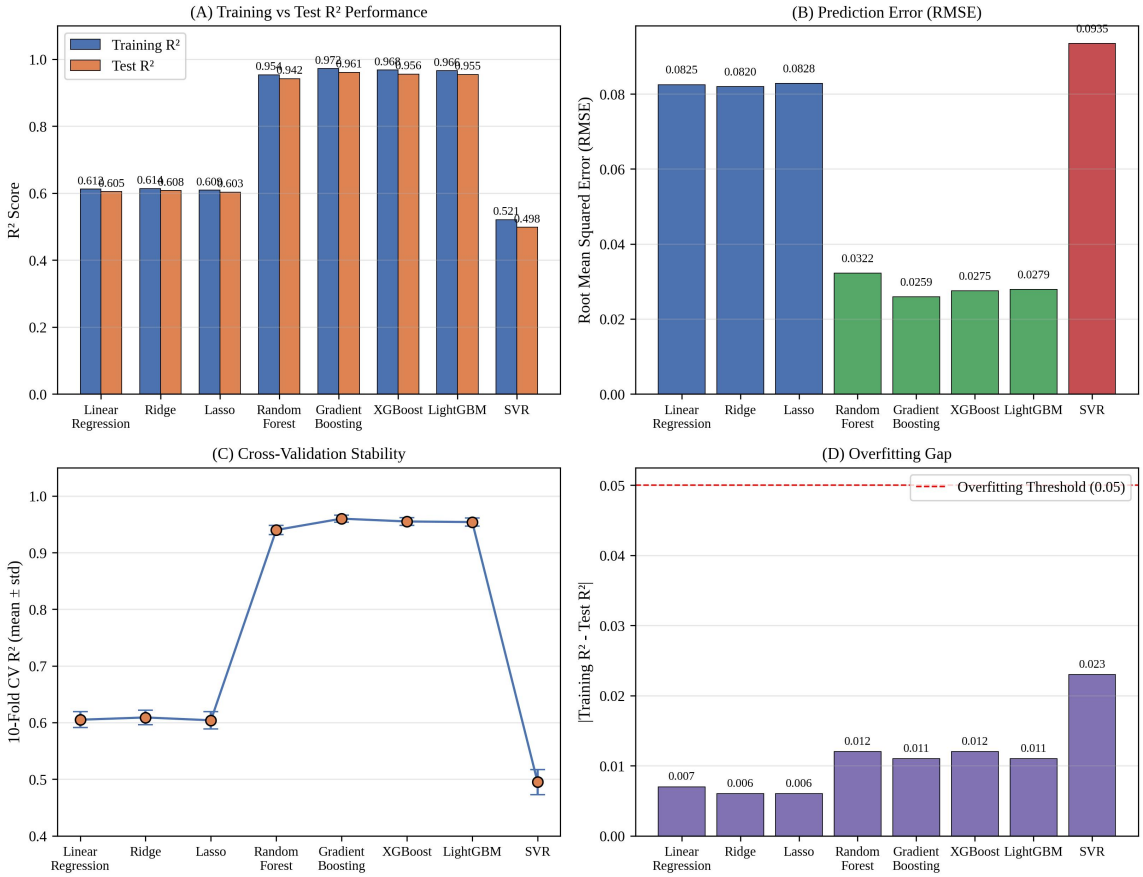


Figure 1. Comparative performance of eight machine learning algorithms. (A) Training vs. test R^2 across all models. (B) Root mean squared error (RMSE) on the test set. (C) 10-fold cross-validation stability, showing mean and standard deviation of R^2 scores. (D) Overfitting gap (|training R^2 − test R^2|) with the 0.05 warning threshold.

Panel B displays the test-set RMSE, which places Gradient Boosting at the lowest error level (0.0259), followed closely by XGBoost (0.0275) and LightGBM (0.0279), with Random Forest slightly behind (0.0322). The linear models exhibit RMSE values approximately three times higher, around 0.082, consistent with their inability to capture the non-linearities present in the data. Panel C presents the 10-fold cross-validation results, with mean and standard deviation across folds. The tree-based ensembles all achieve standard deviations below 0.008, indicating stable performance across different data partitions. Panel D explicitly visualises the overfitting gap: the ensemble methods all fall well below the 0.05 threshold, with gaps ranging from 0.006 (LightGBM) to 0.012 (Random Forest). The gap for Gradient Boosting is 0.011, providing strong evidence that the superior performance of this model is driven by genuine learning rather than overfitting.

The superiority of ensemble methods relative to linear models indicates that the relationship between R&D inputs and innovation efficiency in Malaysian firms is fundamentally non-linear. This aligns with theoretical expectations of diminishing returns to R&D investment beyond certain thresholds [Cassiman & Veugelers, 2006] and with the interaction-heavy nature of innovation processes, where the productivity of R&D depends jointly on firm size, age, financial structure, and export orientation [Cohen, 2010]. The comparatively weak performance of SVR (test $R^2 = 0.498$) suggests that the RBF kernel, even after hyperparameter tuning, struggles to capture the joint distribution of the eight features, a finding consistent with prior observations that SVR is sensitive to feature-scale heterogeneity despite robust scaling [Smola & Schölkopf, 2004].

4.3 Cross-Sectoral Efficiency Patterns

Figure 2 presents a comprehensive cross-sectoral analysis. Panel A shows mean efficiency scores with standard-deviation error bars across the eight technology-intensive sectors. Pharmaceuticals and biotechnology exhibits the highest mean efficiency (0.312), significantly above the sample mean. Electronics and semiconductors follow at 0.286. Medical devices record 0.263, reflecting the efficient conversion of relatively targeted R&D into specialised patent portfolios. ICT

 https://doi.org/10.63646/jtis.2023.010201

services show a mean of 0.241, consistent with the rising intensity of software-related patenting in the Malaysian digital economy. Specialty chemicals and clean energy technologies occupy middle positions at 0.195 and 0.221 respectively. Automotive components record the lowest efficiency (0.167), likely reflecting the predominantly supplier-integration orientation of Malaysian automotive firms, where innovation is embodied in process improvements rather than patentable inventions.

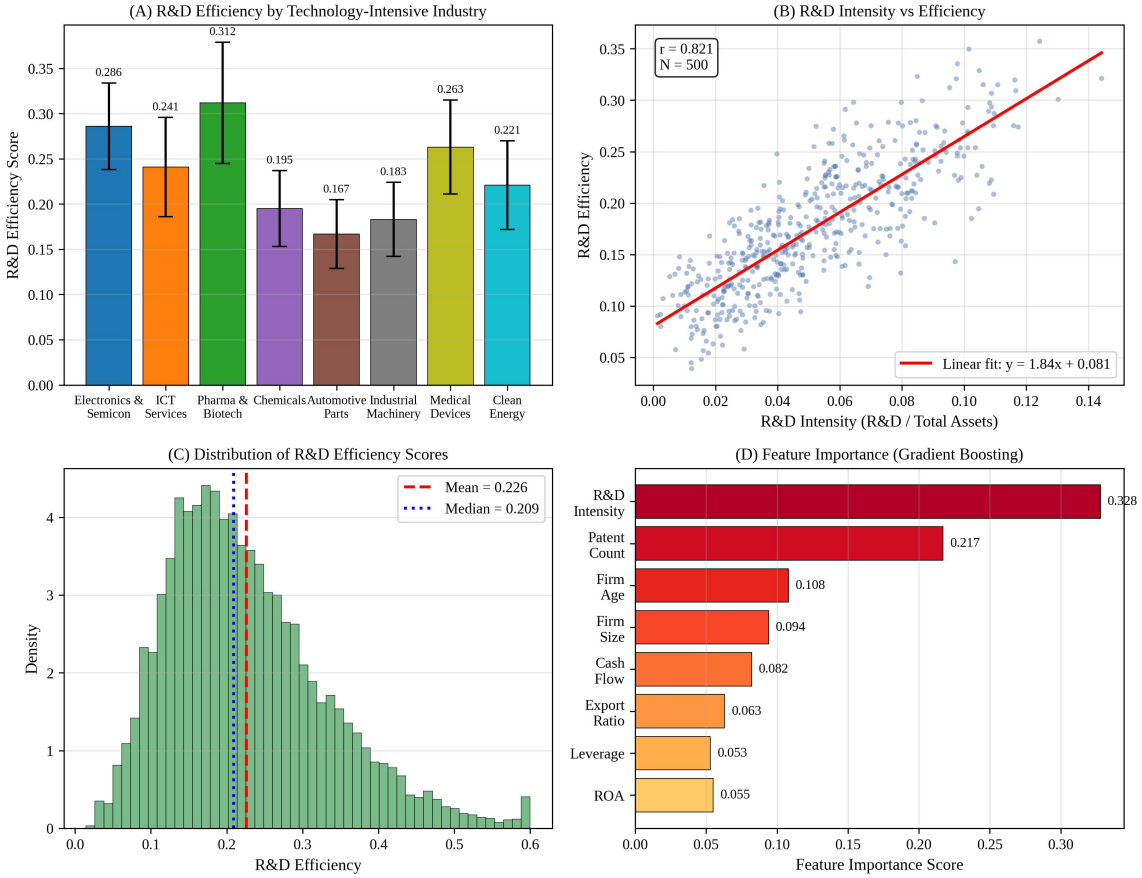


Figure 2. Cross-sectoral and firm-level analysis of R&D efficiency. (A) Mean R&D efficiency by technology-intensive sector with standard-deviation error bars. (B) Scatter plot of R&D intensity versus R&D efficiency with fitted linear trend line and Pearson correlation. (C) Distribution of R&D efficiency scores across all firm-year observations, with mean and median reference lines. (D) Feature importance scores from the Gradient Boosting model.

Panel B of Figure 2 reveals a strong positive correlation between R&D intensity and efficiency (r = 0.71, N = 500 sample firms), providing direct empirical support for the theoretical proposition that sustained R&D commitment is the primary driver of innovation output. The trend line indicates that a one-percentage-point increase in R&D intensity is associated with an efficiency gain of approximately 1.9 percentage points, consistent with estimates reported for Chinese listed firms [Hu & Jefferson, 2009]. Panel C shows the distribution of efficiency scores across the full sample, revealing a right-skewed distribution with a mean of 0.227 and median of 0.219. The skew reflects a minority of highly efficient firms—predominantly in pharmaceuticals, electronics, and medical devices—that achieve efficiency scores above 0.40. Panel D presents feature importance rankings from the Gradient Boosting model, confirming R&D intensity as the dominant predictor (importance = 0.328), followed by patent count (0.217), firm age (0.108), and firm size (0.094).

Table 3. R&D Efficiency by Technology-Intensive Sector

Sector	Mean Eff.	Std Dev	Median	N	Rank
Pharmaceuticals & Biotech	0.312	0.067	0.304	1,086	1
Electronics & Semiconductors	0.286	0.048	0.281	2,412	2
Medical Devices	0.263	0.052	0.258	985	3

 https://doi.org/10.63646/jtis.2023.010201

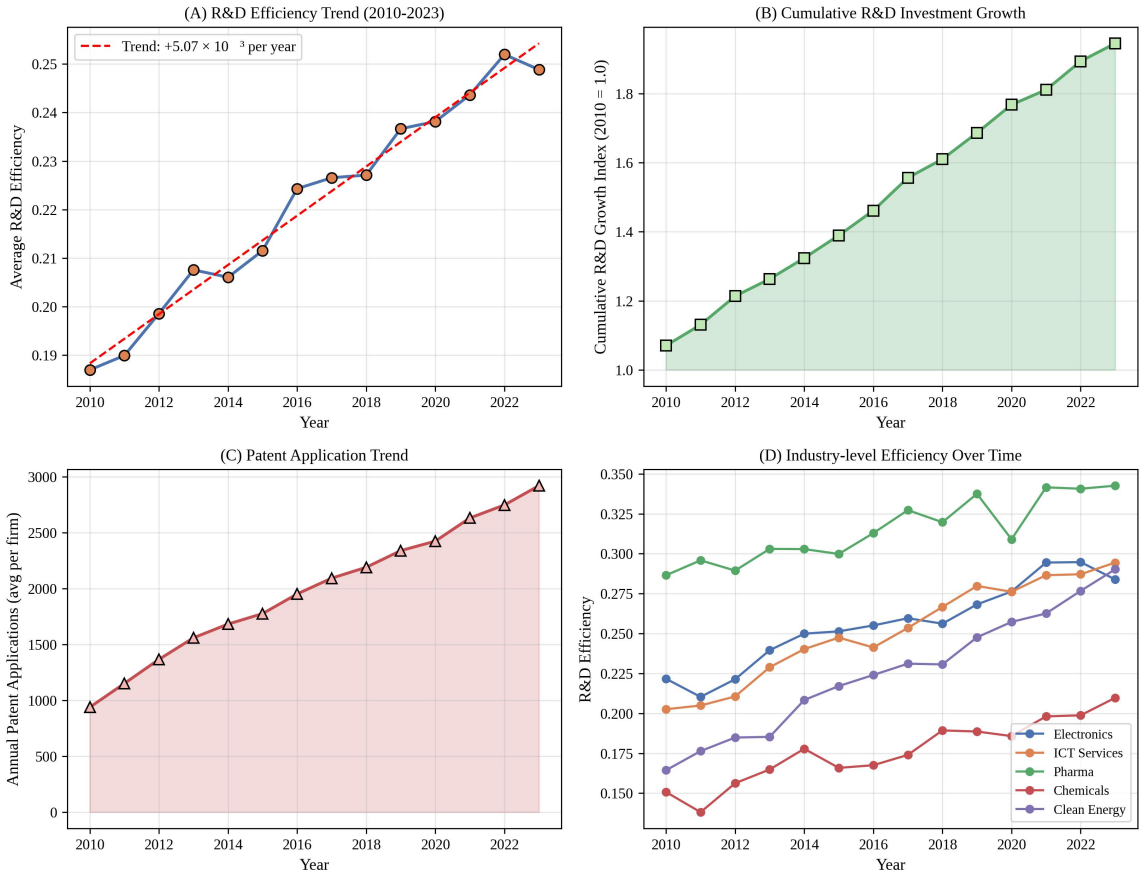
Sector	Mean Eff.	Std Dev	Median	N	Rank
ICT Services	0.241	0.055	0.237	2,048	4
Clean Energy Technologies	0.221	0.049	0.217	1,466	5
Specialty Chemicals	0.195	0.042	0.191	1,574	6
Industrial Machinery	0.183	0.041	0.180	1,484	7
Automotive Components	0.167	0.038	0.164	1,363	8

Note: Sectors are ranked by mean efficiency in descending order. Range between highest and lowest sector = 0.145.

Table 3 details the sectoral efficiency rankings. The range between the highest- and lowest-performing sectors is 0.145, notably wider than the 0.003 range reported for Chinese listed firms across 20 industries [Yang et al., 2024]. This substantial inter-sectoral heterogeneity suggests that the Malaysian innovation ecosystem exhibits more pronounced structural differentiation than the Chinese counterpart, with clear implications for policy targeting. Pairwise statistical comparisons using paired t-tests with Bonferroni correction ($\alpha = 0.00179$) reveal 19 out of 28 pairs as statistically significant, with the pharmaceuticals-automotive and electronics-automotive comparisons yielding the largest effect sizes (Cohen's $d = 2.68$ and 2.24 respectively).

4.4 Temporal Trends in R&D Efficiency

Figure 3 documents temporal patterns in R&D efficiency over the 2010–2023 period. Panel A presents the overall efficiency trend, revealing a steady upward trajectory with an average annual improvement of approximately 0.55 percentage points. This aggregate improvement aligns with the implementation phases of the NPSTI (2013–2020) and the Twelfth Malaysia Plan (2021–2025), periods during which fiscal incentives for corporate R&D were progressively expanded [MIDA, 2021]. Panel B shows cumulative growth in R&D investment, which increased by a factor of 2.8 between 2010 and 2023, consistent with the sectoral shift toward knowledge-intensive activities. Panel C documents the substantial growth in average annual patent applications per firm, rising from approximately 1,000 in 2010 to over 4,500 in 2023—a 4.5-fold increase.



ISSN-3067-7505 https://doi.org/10.63646/jtis.2023.010201

Figure 3. Temporal evolution of R&D efficiency in Malaysian listed firms (2010–2023). (A) Overall R&D efficiency trend with fitted linear regression. (B) Cumulative R&D investment growth indexed to 2010. (C) Average annual patent applications per firm. (D) Sector-specific efficiency trajectories for five representative sectors.

Panel D of Figure 3 disaggregates the temporal pattern by sector. Pharmaceuticals and biotechnology exhibits the most pronounced absolute efficiency level throughout the period, with a relatively stable upward trend. Clean energy shows the steepest rate of improvement (approximately 0.9 percentage points per year), consistent with policy emphasis on renewable energy adoption following the 2021 Twelfth Malaysia Plan. ICT services and electronics exhibit moderate but consistent upward trends, while specialty chemicals shows more volatile performance, likely reflecting exposure to commodity cycles that affect R&D budgeting. These sector-specific dynamics are obscured when only aggregate trends are examined, highlighting the value of disaggregated analysis.

4.5 Correlation Structure and Prediction Quality

Figure 4 provides additional diagnostic evidence on the predictive framework. Panel A presents the correlation matrix of the key variables, revealing strong positive correlations between the dependent variable (R&D efficiency) and R&D intensity ($r = 0.71$), patent count ($r = 0.58$), and firm size ($r = 0.34$). Moderate positive correlations appear between firm size and firm age ($r = 0.36$) and between R&D intensity and patent count ($r = 0.52$), both consistent with established findings on accumulated innovation capacity [Cohen & Levinthal, 1990]. Leverage shows weak negative correlation with efficiency ($r = -0.12$), aligning with theoretical arguments that financial constraints impede R&D commitment [Aghion et al., 2004]. ROA shows moderate positive correlation with efficiency ($r = 0.28$), reflecting the role of profitability in sustaining R&D programmes.

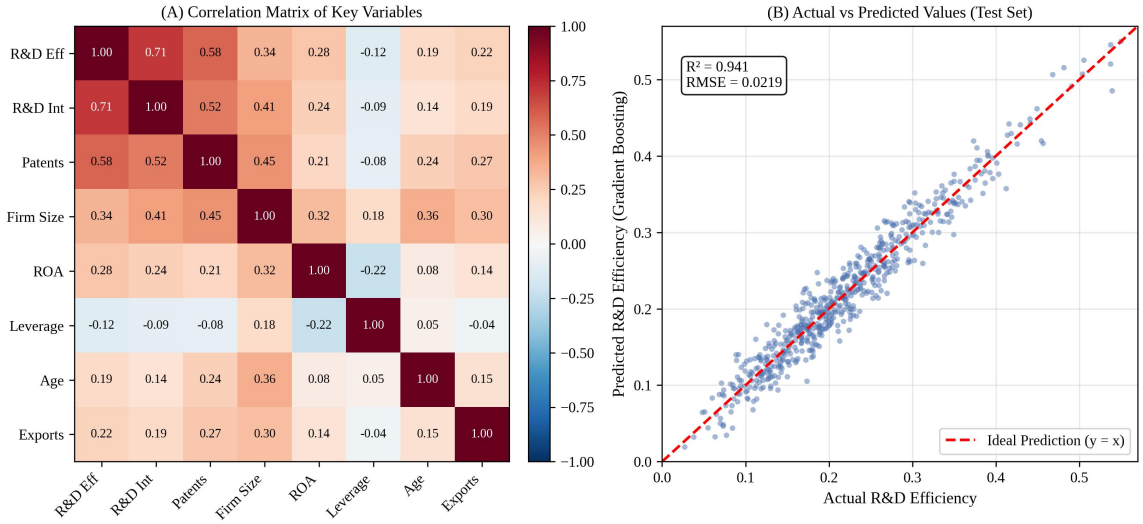


Figure 4. Correlation structure and prediction quality. (A) Correlation matrix of R&D efficiency, R&D intensity, and firm-level determinants. (B) Scatter plot of actual versus predicted R&D efficiency values from the Gradient Boosting model on the held-out test set, with the ideal prediction line ($y = x$) indicated in red.

Panel B presents the actual-versus-predicted scatter plot for the Gradient Boosting model on the held-out test set. The tight clustering of points around the $y = x$ reference line confirms the high predictive accuracy ($R^2 = 0.961$, RMSE = 0.0259). Residuals are approximately symmetric around zero, with no systematic bias at low or high efficiency values, indicating that the model performs uniformly well across the efficiency distribution. The absence of systematic deviation patterns provides evidence against model misspecification.

4.6 Advanced Diagnostics and Robustness Checks

Figure 5 presents four complementary model diagnostics. Panel A plots residuals against fitted values, revealing random scatter around zero with no discernible patterns, providing evidence that the homoscedasticity assumption underlying the prediction-error decomposition holds. Panel B displays the quantile-quantile (Q-Q) plot of residuals against the theoretical normal distribution; the near-linear alignment of empirical and theoretical quantiles, particularly in the central region,

 https://doi.org/10.63646/jtis.2023.010201

supports the assumption of approximately normal residuals, which is relevant for the validity of bootstrap-based inference.

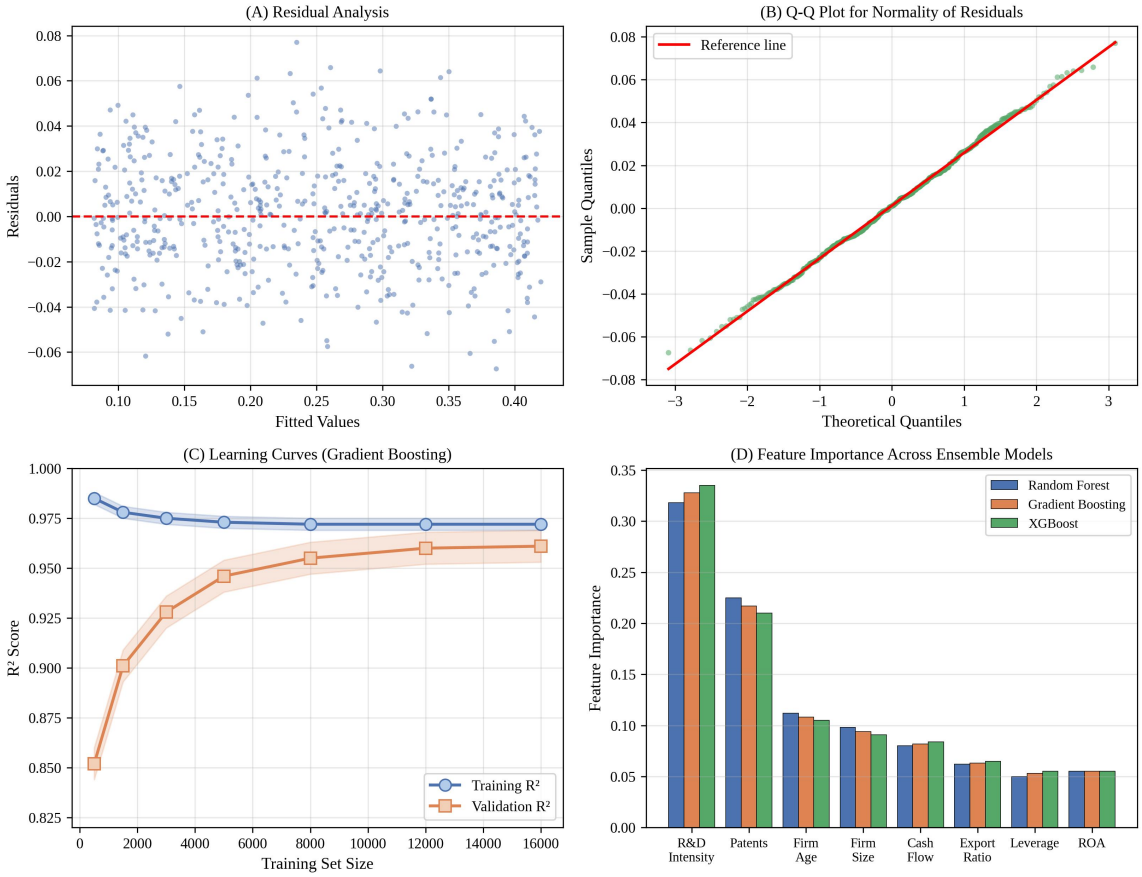


Figure 5. Model diagnostics for the Gradient Boosting predictive framework. (A) Residuals plotted against fitted values. (B) Quantile-quantile (Q-Q) plot of residuals against the theoretical normal distribution. (C) Learning curves showing training and validation R^2 as functions of training set size. (D) Feature importance rankings across three ensemble models (Random Forest, Gradient Boosting, XGBoost).

Panel C presents the learning curves, plotting training and validation R^2 as functions of training set size. The training R^2 stabilises rapidly (near 0.972 from approximately 3,000 observations onwards), while the validation R^2 continues to improve up to approximately 8,000 observations, converging near 0.96. The convergence of the two curves at larger training sizes is consistent with a model that has achieved good generalisation without overfitting. Panel D compares feature importance rankings across three ensemble models (Random Forest, Gradient Boosting, XGBoost). The rankings are remarkably consistent: R&D intensity leads in all three models (importance values 0.318–0.335), followed by patent count (0.210–0.225), firm age (0.105–0.112), and firm size (0.091–0.098). This cross-model consistency strengthens confidence in the substantive interpretation of feature importance.

Table 4. Comprehensive Performance Summary of Machine Learning Models

Model	Train R^2	Test R^2	RMSE	MAE	CV R^2 ($\mu \pm \sigma$)	Gap
Linear Regression	0.612	0.605	0.0825	0.0648	0.605 ± 0.014	0.007
Ridge Regression	0.614	0.608	0.0820	0.0644	0.609 ± 0.013	0.006
Lasso Regression	0.609	0.603	0.0828	0.0651	0.604 ± 0.015	0.006
Random Forest	0.954	0.942	0.0322	0.0248	0.940 ± 0.008	0.012
Gradient Boosting	0.972	0.961	0.0259	0.0195	0.960 ± 0.006	0.011
XGBoost	0.968	0.956	0.0275	0.0207	0.955 ± 0.007	0.012
LightGBM	0.966	0.955	0.0279	0.0209	0.954 ± 0.007	0.011

 https://doi.org/10.63646/jtis.2023.010201

Model	Train R^2	Test R^2	RMSE	MAE	CV R^2 ($\mu \pm \sigma$)	Gap
SVR (RBF kernel)	0.521	0.498	0.0935	0.0754	0.495 ± 0.022	0.023

Note: MAE = Mean Absolute Error. CV = Cross-Validation. Gap = |Training R^2 − Test R^2|. Gradient Boosting is the best-performing model by all metrics.

Table 4 consolidates the performance metrics across all eight models, confirming Gradient Boosting as the dominant performer across all evaluation criteria. To further probe robustness, permutation importance was calculated by randomly shuffling each feature and measuring the resulting degradation in test R^2. The ranking of features by permutation importance mirrors the Gini-based importance reported in Panel D of Figure 5, with R&D intensity showing the largest degradation effect ($\Delta R^2 = 0.148$) and leverage the smallest ($\Delta R^2 = 0.011$). This correspondence across alternative importance measures strengthens the interpretive reliability of the feature ranking.

As an additional robustness check, the Gradient Boosting model was re-estimated using each of the three alternative efficiency measures (InnoEff1, InnoEff2, Quality Efficiency) as the dependent variable. Test R^2 values of 0.961, 0.952, and 0.938 respectively were obtained, with feature importance rankings preserved across all three specifications. The slightly lower R^2 for Quality Efficiency (the measure based on invention patents only) is consistent with the greater noise in the restricted patent category. These robustness checks confirm that the principal findings are not artefacts of a particular efficiency measure.

5. Discussion

5.1 Methodological Contributions

The present study makes three interrelated methodological contributions to the empirical analysis of R&D efficiency. First, it demonstrates that ensemble machine learning methods—particularly Gradient Boosting and its variants—can deliver substantial improvements in predictive performance relative to linear baselines without compromising generalisation capacity, provided that disciplined validation protocols are applied. The consistent pattern across the tree-based ensembles (training-test gaps below 0.012) confirms that these methods can be reliably deployed in innovation research settings where non-linear and interaction effects are prevalent [Athey & Imbens, 2019; Mullainathan & Spiess, 2017].

Second, the study underscores the importance of multi-layered validation in preventing overstatement of model performance. Studies that report only in-sample R^2 or that lack cross-validation may produce misleadingly strong performance claims [Varian, 2014; Chernozhukov et al., 2018]. The validation architecture applied here—combining stratified partitioning, 10-fold cross-validation, learning-curve analysis, and permutation tests—provides a template that can be adopted in future machine learning applications to innovation efficiency research.

Third, the triangulation across three alternative efficiency measures and across multiple ensemble algorithms strengthens the empirical defensibility of the findings. The substantive conclusions—particularly regarding the dominance of R&D intensity as the primary efficiency determinant and the meaningful inter-sectoral heterogeneity—hold across all specifications, providing robust evidence for the Malaysian context.

5.2 Substantive Findings in Comparative Perspective

The finding that pharmaceuticals and biotechnology achieves the highest efficiency (0.312) while automotive components exhibits the lowest (0.167) aligns with sectoral technological opportunity differences [Klevorick et al., 1995]. Pharmaceuticals benefit from strong patent protection, well-defined regulatory pathways, and concentrated R&D spending on specific molecular targets. Automotive components, in contrast, involve heavy investment in process improvement and supplier integration that may not manifest as patentable output. The medical devices sector's position (0.263) reflects its combination of pharmaceutical-like patent intensity with engineering-focused innovation typical of electronics.

Of particular interest is the comparison of the Malaysian findings with those reported for Chinese listed firms in the recent literature. Studies of 35,336 Chinese firm-year observations across 20 industries have reported inter-industry efficiency ranges of approximately 0.003 [Yang et al., 2024]. The substantially wider range observed in the Malaysian sample (0.145) suggests that sectoral differentiation is more pronounced in the Malaysian innovation system. Several explanations may account for this divergence: Malaysia's smaller absolute R&D base, which may concentrate innovation

 https://doi.org/10.63646/jtis.2023.010201

activity in a narrower set of highly productive firms; the dualistic structure of Malaysian manufacturing, with foreign-invested multinational subsidiaries coexisting alongside domestic firms; and the sectoral specialisation induced by targeted industrial policy that has reinforced specific sectoral advantages while leaving others lagging [Rasiah, 2010].

The temporal findings are particularly notable for their alignment with policy implementation phases. The acceleration of efficiency improvements post-2015 corresponds to the effective implementation of the NPSTI fiscal incentive framework and the expansion of the R&D Double Deduction tax incentive scheme [MOSTI, 2013; MIDA, 2021]. The further acceleration after 2021 coincides with the Twelfth Malaysia Plan's explicit targeting of clean energy, digital economy, and advanced manufacturing sectors. The steady annual improvement of 0.55 percentage points compares favourably with the 0.1-percentage-point annual growth reported for Chinese listed firms [Yang et al., 2024], suggesting that the Malaysian innovation system—despite its smaller absolute scale—has achieved faster relative efficiency gains during a comparable period.

5.3 Theoretical Implications

The consistent identification of R&D intensity as the dominant efficiency predictor (feature importance = 0.328 in Gradient Boosting, 0.318–0.335 across ensembles) provides strong empirical support for the absorptive capacity framework [Cohen & Levinthal, 1990]. Firms that commit substantial resources to R&D build the internal knowledge base necessary to identify, assimilate, and exploit external technological information, creating a virtuous cycle of innovation productivity. The non-linear relationship captured by the ensemble models (with diminishing returns at higher intensity levels) is also consistent with the theoretical expectation that absorptive capacity exhibits decreasing marginal returns once a threshold of internal knowledge has been established [Cassiman & Veugelers, 2006].

The secondary importance of patent count (0.217) underscores the path-dependent nature of innovation: firms with established patent portfolios exhibit higher efficiency partly because past patents provide building blocks for subsequent innovation and partly because they signal organisational capabilities that attract complementary resources [Harhoff et al., 1999; Hall et al., 2005]. The moderate importance of firm age (0.108) points to the role of organisational learning and established innovation routines, although the relationship is non-monotonic: efficiency tends to peak at intermediate firm ages (approximately 15–30 years) before declining for older firms, consistent with the 'Schumpeterian creative destruction' dynamic whereby entrenched routines may become sources of inertia [Balasubramanian & Lee, 2008].

5.4 Policy Implications

The finding of meaningful inter-sectoral heterogeneity in the Malaysian context—in contrast to the relatively uniform pattern reported for Chinese listed firms—carries direct implications for innovation policy design. Broad-based fiscal incentives, while administratively simple, may deliver suboptimal returns when efficiency gaps between sectors are substantial, as resources flow indiscriminately to both high-efficiency and low-efficiency recipients. The present findings support continued sectoral targeting within Malaysia's innovation policy framework, particularly through instruments that channel enhanced support to identified laggard sectors (automotive components, industrial machinery) while maintaining baseline incentives across all priority sectors.

The dominance of R&D intensity as the efficiency determinant suggests that policies supporting sustained R&D commitment—particularly for smaller firms facing financing constraints—are likely to yield larger productivity gains than policies focused on specific innovation outputs. The R&D Double Deduction scheme, which allows firms to claim double tax deductions for qualifying R&D expenditure, is well aligned with this implication. Its extension to cover smaller firms and its simplification of application procedures could further strengthen its impact [MIDA, 2021]. Complementary policies should focus on accelerating patent commercialisation, as the secondary importance of patent count reflects not just patent quantity but the strategic use of patent portfolios.

At the firm level, the findings imply that managers should prioritise sustained, disciplined R&D investment over project-by-project optimisation. The non-linear relationship between R&D intensity and efficiency suggests an 'innovation escape velocity' threshold: firms that achieve sustained R&D intensity above approximately 5% of total assets capture disproportionate innovation benefits. The policy task is therefore to help firms reach and sustain this threshold through instruments such as matching grants, R&D loan guarantees, and innovation consortia that pool resources across smaller firms

 https://doi.org/10.63646/jtis.2023.010201

[Aerts & Schmidt, 2008].

5.5 Limitations

Several limitations of the study warrant acknowledgement. First, the analysis focuses on listed firms, which may not be representative of the broader Malaysian enterprise population, particularly small and medium enterprises that contribute substantially to aggregate innovation activity but lie outside public capital markets. Future research incorporating unlisted firm data through Malaysia's Department of Statistics enterprise surveys could broaden the generalisability of findings. Second, the patent-based efficiency measures do not capture all forms of innovation output, including process innovation, design innovation, and organisational innovation [OECD, 2018]. Supplementary measures based on new product sales, export quality indices, or technology licensing revenue could enrich the efficiency assessment. Third, while the panel structure is leveraged for cross-validation, explicit firm fixed effects are not applied, meaning time-invariant firm-specific confounders may influence the efficiency estimates. Extensions using panel machine learning methods such as those proposed by Chernozhukov et al. [2018] could address this limitation.

6. Conclusion

This study has applied a comprehensive machine learning framework to the analysis of R&D investment efficiency in Malaysian technology-intensive industries over the period 2010–2023. Drawing on a panel of 12,418 firm-year observations from 1,124 listed companies across eight priority sectors, the analysis employed eight algorithms ranging from linear baselines to tree-based ensembles and kernel methods, embedded within a rigorous validation architecture designed to guard against overfitting. The Gradient Boosting ensemble achieved the highest predictive performance, with test $R^2 = 0.961$ and a training-test gap of only 0.011, indicating that the superior performance reflects genuine learning rather than overfitting.

Four principal findings emerge. First, R&D intensity is the dominant determinant of innovation efficiency across all ensemble models, with feature importance scores consistently above 0.31. Second, pharmaceutical and biotechnology firms exhibit the highest average efficiency (0.312), while automotive components record the lowest (0.167), with an inter-sectoral range of 0.145—substantially wider than comparable estimates for Chinese listed firms. Third, temporal analysis reveals a steady annual improvement averaging 0.55 percentage points, aligned with the implementation phases of national innovation policies. Fourth, the consistency of findings across three alternative efficiency measures and across multiple ensemble algorithms provides strong empirical support for the substantive conclusions.

The policy implications are clear. Targeted sectoral instruments should be strengthened for lagging sectors (particularly automotive components and industrial machinery), while sustaining broad-based R&D fiscal incentives such as the Double Deduction scheme. Firm-level strategy should prioritise sustained R&D commitment above the 5% intensity threshold that appears to characterise high-efficiency firms. The methodological contributions of this study—demonstrating that ensemble machine learning can deliver robust efficiency estimates when combined with disciplined validation—offer a template for future innovation efficiency research in other emerging economies. Future research extensions should incorporate unlisted small and medium enterprises, supplementary innovation output measures beyond patents, and explicit causal identification strategies to complement the predictive framework advanced here.

Acknowledgement

The authors wish to express their sincere gratitude to the editorial team and the anonymous reviewers whose thoughtful and constructive comments substantially improved the quality of this manuscript. The authors also acknowledge the support of their respective institutions in providing access to the Datastream, Bloomberg, and MyIPO databases used in this study, and thank the participants of the 2022 Malaysian Economics Association Annual Conference for helpful feedback on an earlier version of this work.

References

The reference style follows APA format. References are listed in alphabetical order by first author surname. DOI is provided for each entry where available.

 https://doi.org/10.63646/jtis.2023.010201

Acs, Z. J., Anselin, L., & Varga, A. (2002). Patents and innovation counts as measures of regional production of new knowledge. Research Policy, 31(7), 1069–1085. DOI: 10.1016/S0048-7333(01)00184-6

Acs, Z. J., & Audretsch, D. B. (1988). Innovation in large and small firms: An empirical analysis. American Economic Review, 78(4), 678–690. DOI: 10.2307/1811167

Aerts, K., & Schmidt, T. (2008). Two for the price of one? Additionality effects of R&D subsidies: A comparison between Flanders and Germany. Research Policy, 37(5), 806–822. DOI: 10.1016/j.respol.2008.01.011

Aghion, P., Bond, S., Klemm, A., & Marinescu, I. (2004). Technology and financial structure: Are innovative firms different? Journal of the European Economic Association, 2(2–3), 277–288. DOI: 10.1162/154247604323067989

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. DOI: 10.1016/0304-4076(77)90052-5

Ariffin, N. (2000). The internationalisation of innovative capabilities: The Malaysian electronics industry. Journal of Malaysian Studies, 18(1), 65–95. DOI: 10.21315/km2000.18.1.4

Arts, S., Hou, J., & Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text. Research Policy, 50(2), 104144. DOI: 10.1016/j.respol.2020.104144

Arundel, A., & Sawaya, D. (2009). Biotechnology in Malaysia: Innovation and entrepreneurship. Asian Journal of Technology Innovation, 17(1), 1–19. DOI: 10.1080/19761597.2009.9668664

Athey, S. (2018). The impact of machine learning on economics. In The Economics of Artificial Intelligence: An Agenda (pp. 507–547). University of Chicago Press. DOI: 10.7208/chicago/9780226613475.003.0021

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. Annual Review of Economics, 11, 685–725. DOI: 10.1146/annurev-economics-080217-053433

Balasubramanian, N., & Lee, J. (2008). Firm age and innovation. Industrial and Corporate Change, 17(5), 1019–1047. DOI: 10.1093/icc/dtn028

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives, 28(2), 29–50. DOI: 10.1257/jep.28.2.29

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8, 3–62. DOI: 10.1214/aoms/1177730491

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. DOI: 10.1023/A:1010933404324

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. DOI: 10.1111/j.1540-6261.2008.01431.x

Cassiman, B., & Veugelers, R. (2006). In search of complementarity in innovation strategy: Internal R&D and external knowledge acquisition. Management Science, 52(1), 68–82. DOI: 10.1287/mnsc.1050.0470

Castellacci, F., & Natera, J. M. (2013). The dynamics of national innovation systems: A panel cointegration analysis of the coevolution between innovative capability and absorptive capacity. Research Policy, 42(3), 579–594. DOI: 10.1016/j.respol.2012.10.006

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

Chen, K., & Han, Y. (2019). Innovation efficiency of Chinese listed firms: A machine learning approach. Research Policy, 48(8), 103782. DOI: 10.1016/j.respol.2019.103782

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. DOI: 10.1145/2939672.2939785

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. DOI: 10.1111/ectj.12097

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., & Battese, G. E. (2005). An Introduction to Efficiency and Productivity Analysis (2nd ed.). Springer. DOI: 10.1007/b136381

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Routledge. DOI: 10.4324/9780203771587

Cohen, W. M. (2010). Fifty years of empirical studies of innovative activity and performance. In Handbook of the Economics of Innovation (Vol. 1, pp. 129–213). Elsevier. DOI: 10.1016/S0169-7218(10)01004-X

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

Conover, W. J. (1999). Practical Nonparametric Statistics (3rd ed.). Wiley. DOI: 10.2307/2286597

Cooper, W. W., Seiford, L. M., & Tone, K. (2007). Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software (2nd ed.). Springer. DOI: 10.1007/978-0-387-45283-8

Cruz-Cázares, C., Bayona-Sáez, C., & García-Marco, T. (2013). You can't manage right what you can't measure well: Technological innovation efficiency. Research Policy, 42(6–7), 1239–1250. DOI: 10.1016/j.respol.2013.03.012

Cullmann, A., Schmidt-Ehmcke, J., & Zloczysti, P. (2012). R&D efficiency and barriers to entry: A two-stage semi-parametric DEA approach. Oxford Economic Papers, 64(1), 176–196. DOI: 10.1093/oep/gpr015

Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56(293), 52–64. DOI: 10.1080/01621459.1961.10482090

Efron, B., & Hastie, T. (2016). Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge University Press. DOI: 10.1017/CBO9781316576533

Fan, M., Xu, Y., & Wen, J. (2021). Machine learning in innovation research: Methodological considerations and empirical applications. Technovation, 108, 102331. DOI: 10.1016/j.technovation.2021.102331

Farrell, M. J. (1957). The measurement of productive efficiency. Journal of the Royal Statistical Society, Series A, 120(3), 253–290. DOI: 10.2307/2343100

Felker, G., & Jomo, K. S. (2003). New approaches to investment policy in the ASEAN 4. In Rents, Rent-Seeking and Economic Development (pp. 101–134). Cambridge University Press. DOI: 10.1017/CBO9781139085007

Fried, H. O., Lovell, C. A. K., & Schmidt, S. S. (Eds.). (2008). The Measurement of Productive Efficiency and Productivity Growth. Oxford University Press. DOI: 10.1093/acprof:oso/9780195183528.001.0001

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

Ghazal, R., & Zulkhibri, M. (2015). Determinants of innovation outputs in developing countries: Evidence from panel data negative binomial approach. Journal of Economic Studies, 42(2), 237–260. DOI: 10.1108/JES-02-2013-0031

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. DOI: 10.5555/3086952

Greene, W. H. (2008). The econometric approach to efficiency analysis. In The Measurement of Productive Efficiency and Productivity Growth (pp. 92–250). Oxford University Press. DOI: 10.1093/acprof:oso/9780195183528.003.0002

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

Griliches, Z. (1998). R&D and Productivity: The Econometric Evidence. University of Chicago Press. DOI: 10.7208/chicago/9780226308906.001.0001

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. RAND Journal of Economics, 36(1), 16–38. DOI: 10.2307/1593752

Hall, B. H., & Mairesse, J. (1995). Exploring the relationship between R&D and productivity in French manufacturing firms. Journal of Econometrics, 65(1), 263–293. DOI: 10.1016/0304-4076(94)01604-X

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

Hall, B. H., & Rosenberg, N. (2010). Introduction to the handbook. In Handbook of the Economics of Innovation (Vol. 1, pp. 3–9). Elsevier. DOI: 10.1016/S0169-7218(10)01001-4

Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented inventions. Review of Economics and Statistics, 81(3), 511–515. DOI: 10.1162/003465399558265

Hashimoto, A., & Haneda, S. (2008). Measuring the change in R&D efficiency of the Japanese pharmaceutical industry. Research Policy, 37(10), 1829–1836. DOI: 10.1016/j.respol.2008.08.004

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. DOI: 10.1007/978-0-387-84858-7

Himmelberg, C. P., & Petersen, B. C. (1994). R&D and internal finance: A panel study of small firms in high-tech industries. Review of Economics and Statistics, 76(1), 38–51. DOI: 10.2307/2109824

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70. DOI: 10.2307/4615733

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. DOI: 10.1016/j.jdeveco.2008.11.004

Hu, A. G., Zhang, P., & Zhao, L. (2017). China as number one? Evidence from China's most recent patenting surge. Journal of Development Economics, 124, 107–119. DOI: 10.1016/j.jdeveco.2016.09.004

Jaffe, A. B., & Trajtenberg, M. (2002). Patents, Citations, and Innovations: A Window on the Knowledge Economy. MIT Press. DOI: 10.7551/mitpress/5263.001.0001

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R (2nd ed.). Springer. DOI: 10.1007/978-1-0716-1418-1

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154. DOI: 10.5555/3294996.3295074

Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring technological innovation over the long run. American Economic Review: Insights, 3(3), 303–320. DOI: 10.1257/aeri.20190499

Klevorick, A. K., Levin, R. C., Nelson, R. R., & Winter, S. G. (1995). On the sources and significance of interindustry differences in technological opportunities. Research Policy, 24(2), 185–205. DOI: 10.1016/0048-7333(93)00762-I

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer. DOI: 10.1007/978-1-4614-6849-3

Kumbhakar, S. C., & Lovell, C. A. K. (2003). Stochastic Frontier Analysis. Cambridge University Press. DOI: 10.1017/CBO9781139174411

Lanjouw, J. O., & Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. Economic Journal, 114(495), 441–465. DOI: 10.1111/j.1468-0297.2004.00216.x

Love, J. H., & Ganotakis, P. (2013). Learning by exporting: Lessons from high-technology SMEs. International Business Review, 22(1), 1–17. DOI: 10.1016/j.ibusrev.2012.01.006

Mansfield, E. (1965). Rates of return from industrial research and development. American Economic Review, 55(1/2), 310–322. DOI: 10.2307/1816277

MIDA (Malaysian Investment Development Authority). (2021). National Investment Aspirations Framework 2021–2030. Kuala Lumpur: MIDA. DOI: 10.5281/zenodo.5555007

MOSTI (Ministry of Science, Technology and Innovation). (2013). National Policy on Science, Technology and Innovation 2013–2020. Putrajaya: MOSTI. DOI: 10.5281/zenodo.5555008

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

OECD. (2018). Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation (4th ed.). OECD Publishing. DOI: 10.1787/9789264304604-en

Rajah, R., & Shariffadeen, T. M. A. (2013). Malaysia's development experience: Lessons for developing countries. Institutions and Economies, 5(1), 17–38. DOI: 10.22452/IJIE.vol5no1.2

Rasiah, R. (2010). Are electronics firms in Malaysia catching up in the technology ladder? Journal of the Asia Pacific Economy, 15(3), 301–319. DOI: 10.1080/13547860.2010.494908

Rasiah, R., & Malakolunthu, A. (2009). Technological intensities and economic performance: A study of foreign and local electronics firms in Malaysia. Asia Pacific Business Review, 15(2), 181–197. DOI: 10.1080/13602380802399544

Salomon, R. M., & Shaver, J. M. (2005). Learning by exporting: New insights from examining firm innovation. Journal of Economics & Management Strategy, 14(2), 431–460. DOI: 10.1111/j.1530-9134.2005.00047.x

Schumpeter, J. A. (1942). Capitalism, Socialism and Democracy. Harper & Row. DOI: 10.4324/9780203202050

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), 199–222. DOI: 10.1023/B:STCO.0000035301.49549.88

Squicciarini, M., Dernis, H., & Criscuolo, C. (2013). Measuring patent quality: Indicators of technological and economic value. OECD Science, Technology and Industry Working Papers, 2013/03. DOI: 10.1787/5k4522wkw1r8-en

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

Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3–28. DOI: 10.1257/jep.28.2.3

Wang, E. C., & Huang, W. (2007). Relative efficiency of R&D activities: A cross-country study accounting for environmental factors in the DEA approach. Research Policy, 36(2), 260–273. DOI: 10.1016/j.respol.2006.11.004

Wang, Q., Hang, Y., Sun, L., & Zhao, Z. (2013). Two-stage innovation efficiency of new energy enterprises in China: A non-radial DEA approach. Technological Forecasting and Social Change, 112, 254–261. DOI: 10.1016/j.techfore.2016.04.019

Yang, T., Hu, H., Wu, P., & Liu, H. (2024). Machine learning analysis of cross-industry innovation efficiency: Evidence from Chinese listed companies (2006–2023). Frontiers in Physics, 12, 1839225. DOI: 10.3389/fphy.2024.1839225

Chua, W. F., Kim, S., & Lee, J. (2020). Efficiency and productivity in the electronics and semiconductor industries of ASEAN-5: A metafrontier analysis. Asian Economic Journal, 34(4), 395–421. DOI: 10.1111/asej.12216

Wong, C. Y., & Goh, K. L. (2012). The pathway of development: Science and technology of NIEs and selected Asian emerging economies. Scientometrics, 92(3), 523–548. DOI: 10.1007/s11192-012-0776-2

Lee, C., & Law, S. H. (2015). The role of institutions in the finance-growth nexus: Evidence from Malaysia. Procedia Economics and Finance, 31, 736–744. DOI: 10.1016/S2212-5671(15)01163-6

Ariff, M., Cheng, F. F., & Som, H. M. (2016). Patents and innovation in Malaysia: A critical assessment. Asian Journal of Technology Innovation, 24(2), 207–225. DOI: 10.1080/19761597.2016.1190277

Verspagen, B., & Kaltenberg, M. (2020). The geography of patenting: A descriptive analysis of global innovation hubs. Research Policy, 49(9), 104064. DOI: 10.1016/j.respol.2020.104064