Feature Analysis of Patents——Data Mining in the Elevator Industry

Yuhang Guan, Runzheng Wang

School of Computer Science and Technology, Shandong University, Qingdao 266237, China

Abstract

Keywords:

1. Introduction

According to WIPO patent data statistics[1], starting from 2009, the global patent application count has steadily increased, achieving eight consecutive years of growth.The global patent industry has shown a general trend of expansion. This is attributed to the increasing innovation activities, particularly in the fields of technology, healthcare, and biotechnology. Some emerging market countries, such as China and India, have experienced significant growth in the patent domain, becoming important participants in global patent activities.

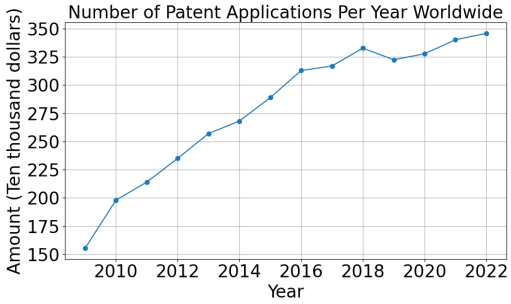


Figure 1: The Global Annual Number of Patent Applications

The implementation of patents requires financing. When patent financing institutions provide loans, they first need to consider various patent indicators to estimate their value. Subsequently, the loan amount is determined based on the assessed value of the patents. In the past, the estimation

[1]Source: https://www.wipo.int/pressroom/en/

of patent value relied on rough data statistics and subjective judgments of patent financing institutions. However, these methods become less effective when dealing with patents with insufficient information or containing excessive abstract information. Irrational loans may lead to financial issues for institutions and have a negative impact on future patent financing. This poses challenges to patent financing.

Currently, in the research on patent value, the focus is primarily on how to construct a patent evaluation indicator system. The current research emphasis is on analyzing patent value using AI models and traditional data analysis methods. This article takes the elevator industry as a case study, characterizes various quantitative and categorical labels of patents through feature extraction and correlation analysis, and constructs a patent evaluation indicator system. Additionally, we use a decision tree classification(DTC[Quinlan, 1986]) model to build a Patent Value Classification Prediction (**PVCP**) model for predicting the value of patents. This prediction aims to provide guidance for patent financing institutions.

Our main contributions are summarized as follows:

1.We have constructed a patent evaluation index system.

2.we have provided patent financing institutions with scientifically reliable patent valuation references.

2. Related work

2.1. Patent Application

Early patent analysis methods were incomplete, with only a few qualitative and quantitative analysis methods emerging. In the field of patent planning, Jeone and Yoon pioneered the use of quantitative patent analysis methods to propose a new technological roadmap[Jeong and Yoon, 2011] based on the patent citation network, providing strategic planning for patent technology. Subsequently, with the improvement of data mining methods[Agrawal et al., 1993], Chaoan and Cuilu combined data mining and statistical analysis to create a visualization-based patent mining analysis framework[Chaoan and Cuilu, 2016] for discovering technological hotspots and patent gaps in smart homes. Progress has also been made in the application of patent informatization in patent retrieval. Comins and Leydesdorff introduced the Patent Citation Spectrum (PCS) method, a fundamental component in constructing the patent landscape for biomedical research and development, aimed at identifying the most groundbreaking patents. Other applications of patent

informatization include patent evaluation, which will be detailed in the following sections.

2.2. Patent Evaluation

The concept of "patent value" was first introduced by Sanders[SANDERS B S, 1958], who noted that only 55% of patents are implemented, and there is a significant difference in the commercial value among patents. Patent value has two dimensions, internal and external. Internal value primarily refers to the technical aspects of a patent, such as technological advancement, technical complexity, and legal aspects like the number of claims, family size, and citation count after grant. External value pertains to a patent's potential marketability and its relevance to patent market value, economic value, novelty, practicality, patent lifespan, and other related factors. Subsequently, We will systematically review the evaluation of patent value from two aspects: evaluation indicators and evaluation methods. Evaluation Indicators. In the initial studies, researchers adopted a single-indicator method for assessing patent value, with citation count being a common choice[Trajtenberg, 1990]. However, this method has faced scrutiny. A research report from the Rand Corporation indicated that the citation of patents does not exhibit a linear relationship with their value, and the positive correlation between the two is not ideal. Other commonly used indicators include patent lifespan[Schankerman and Pakes, 1986], patent grant rate[Griliches, 1998], family size[Neuhäusler and Frietsch, 2013], and number of claims[Llanes and Trento, 2012]. While this method is straightforward, it fails to comprehensively reflect the value of patents. Subsequently, researchers turned to a multi-indicator comprehensive assessment. CHI Research and the National Science Foundation in the United States proposed, for the first time, a comprehensive assessment of patent value for regions and overall enterprises, considering seven classical patent value evaluation indicators, including the quantity of patents, citation count, and technological lifecycle. Harhoff et al.[Harhoff et al., 2003] selected indicators such as patent scope and family size to evaluate patent value. Park[Park and Park, 2004] constructed a patent value evaluation indicator system by selecting multiple relevant indicators involving intrinsic technological features and technology usage. The combination of multiple factors is currently the mainstream method for patent value assessment, reducing the one-sidedness and subjectivity of single-indicator assessments, although it still faces challenges such as redundant indicator settings and unreasonable weightings.

Evaluation Methods. In the early stages, scholars adopted the cost method, which was not proposed by a specific individual but gradually formed through the research of multiple economists and assessment experts in various fields. However, due to its neglect of the impact of time on patent value and the inability to consider market demand and economic principles, this method was gradually phased out. Subsequently, scholars improved and proposed several methods based on the cost method, with the notable European IPScore system among them. However, these methods only qualitatively describe patent value and cannot provide reliable data support.In the 1980s, with the introduction of Partial Least Squares (PLS)[Wold, 1975] into the fields of social science and management research, patent value indicators underwent quantitative analysis. In 2009, Martinez Ruiz Alba and Aluja Banet Tomas proposed the PLS path modeling, connecting variables that determine patent value. They demonstrated excellent performance in correlational analysis, taking the renewable energy sector as an example.Subsequently, the rise of machine learning models brought new approaches to patent value research. Secil Ercan and Gulgun Kayakutlu used the Support Vector Machine(SVM) model[Cortes and Vapnik, 1995] to construct an intelligent classification model in the household appliance patent industry[Ercan and Kayakutlu, 2014], predicting the likelihood of funding and assisting decision-makers in anticipating whether a patent appeal would be accepted.

3. Basic Theory and Research Design

3.1. Basic Theory and Methods

3.1.1. Traditional Methods

Descriptive statistical analysis. The main purpose of this method is to better understand the distribution and characteristics of the data by summarizing and analyzing basic statistical information about the data set. Statistical information includes information such as mean, median, variance, standard deviation, and quantile.

ANOVA. ANOVA is a statistical test used to compare differences in means between three or more groups (or samples). Its main purpose is to determine whether at least one group has a mean that is different from the others.

Pearson correlation coefficient. Pearson's correlation coefficient is a statistical method used to measure the linear relationship between two variables. The method assesses the degree of correlation between these two variables by calculating the correlation coefficient, which has a value between -1 and 1,

reflecting the direction and strength of the correlation. The Pearson correlation coefficient standardizes the covariance by dividing it by the product of the standard deviations of the two variables, resulting in a unit-independent measure. This allows r to quantify the strength and direction of the linear relationship between two variables.

3.1.2. AI Methods

DTC. DTC algorithm is a common supervised learning algorithm for solving classification problems. It is based on the principle of tree structure, where data is partitioned by recursively selecting the best features to construct a decision tree which is used to classify new samples. One of the advantages of the DTC algorithm is its interpretability. The generated decision tree visualizes the features and thresholds on which each decision node is based, making the decision-making process easy to understand. In addition, decision trees are suitable for a variety of data types, including categorical and numerical features. Another advantage is its high computational efficiency when dealing with large datasets.

3.2. Research Design

3.2.1. Patent Data Collection

This study's patent data was provided by manufacturers in the elevator industry. In total, we collected 252,047 records, with each patent information containing 50 indicators.

3.2.2. Feature Selection

Referring to the achievements of previous scholars and the policies outlined in China's "14th Five-Year National Plan for Intellectual Property Protection and Utilization," we selected 15 patent indicators based on their scientific relevance and availability. These indicators include: Industry Chain Position, Primary Technical Branch, Patent Type, Publication Country, Publication Date, Number of Claims, Number of Document Pages, IPC (International Patent Classification), Applicant's Province and City Code, Number of Citations within 3 Years, Number of Citations within 5 Years, Number of Citing Patents, Number of Cited Patents, Patent Validity, and Number of Litigation Cases, Patent Lifespan.

Among these indicators, the categorical ones are Industry Chain Position, Primary Technical Branch, Patent Type, Publication Country, IPC, Applicant's Province and City Code, and Patent Validity. The quantitative indi-

cators include Publication Date, Number of Claims, Number of Document Pages, Number of Citations within 3 Years, Number of Citations within 5 Years, Number of Citing Patents, Number of Cited Patents, and Number of Litigation Cases, Patent Lifespan.

3.2.3. Feature and Correlation Analysis

For categorical indicators, we first calculate the mean, median, variance, and other statistics of patent values within each indicator. We use ANOVA to analyze the differences in patent values among these indicators. For quantitative indicators, we employ Pearson correlation coefficients to analyze the relationships between pairs of indicators. Through these two approaches, we obtain the degree of correlation between each indicator and patent value.

3.2.4. Construction of PVCP Model

We used the aforementioned 15 indicators as model features. Due to the DTC model's strong interpretability and efficient processing of large datasets, we chose it as our training model. To ensure the accuracy of the results, we employed cross-validationBox and Meyer [1986] during the model training process.

4. Experimental Implementation

4.1. Data Preprocessing

After removing data points with missing values for the selected 15 indicators, we obtained a final dataset of 72,037 records. As more than 50% of the data points had missing values for the patent lifespan indicator, we decided to exclude this indicator. In the end, we retained a dataset comprising 72,037 records with 14 indicators.

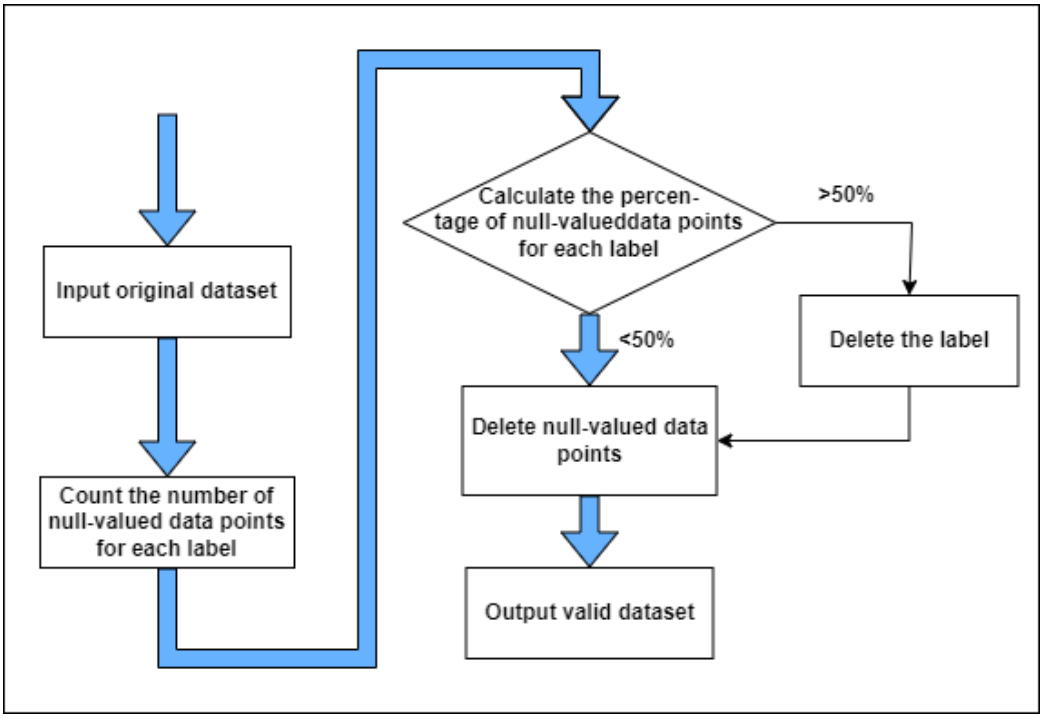
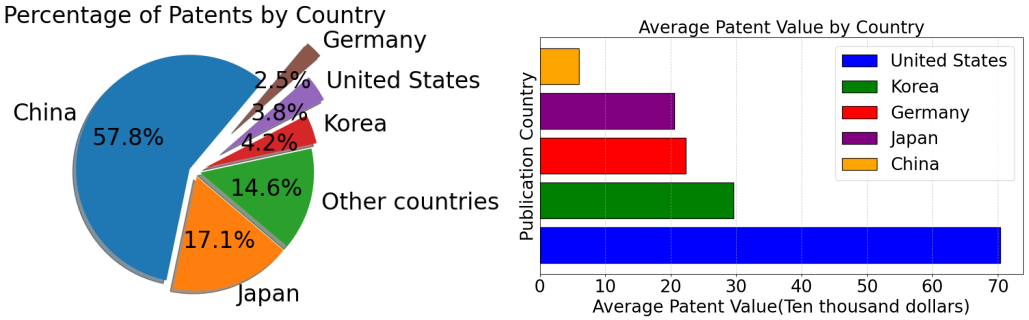


Figure 2: Data preprocessing process

4.2. Global Patent Analysis

1. Publication Country. We have compiled the number of elevator patents for each country. Given the numerous countries, we primarily focused on China, the United States, Japan, South Korea, and Germany for our analysis. The results are shown in Figure 1.



(a) Proportion of Patents by Country

(b) Average patent value by country

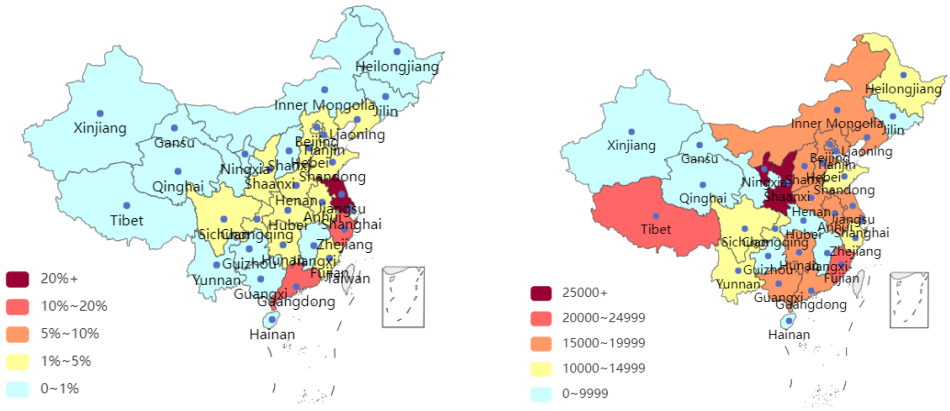
Figure 3: Distribution of the Proportion and Average Value of Patents by Country

From the chart, we can observe that China accounts for 57.8% of the total number of patents in the elevator industry, followed by Japan at 37.1%, South Korea at 4.2%, the United States at 3.8%, and Germany at 2.5%. Other countries collectively make up 14.6%, indicating variations in patent numbers among different nations. Furthermore, the mean patent value ranks as follows, from lowest to highest: China, Japan, Germany, South Korea, and the United States. This reflects the common trend of foreign patents having higher values, primarily due to the higher costs associated with applying for foreign patents compared to domestic ones. It is evident that patent values also vary across different countries. In conclusion, 'Publication Country' is a relevant feature.

4.3. Chinese Patent Analysis

4.3.1. Analysis of Categorical Features

2. Applicant's Province and City Code. We calculated the proportion of patent numbers and the mean patent value for each province in China, as shown in Figure 2.



(a) Proportion of Patents by Province of China (b) Average Patent Value by Province of China

Figure 4: Proportion and Average Value of Patents by Province of China

From the chart, it is evident that coastal provinces account for over 75% of the total patent quantity in the entire country. Additionally, the mean patent value is higher in these coastal provinces as well as in inland provinces that are closer to the sea. This indicates that coastal regions have a certain technological advantage in patents, and it further suggests a correlation between 'Applicant's Province and City Code' and patent value.

3. Publication Date.Publication Date. Publication Date is a temporal attribute that can be analyzed as both a categorical and a quantitative feature. In this section, we will begin by examining its role as a categorical feature, as depicted in Figure 3.

The chart reveals that from 2003 to 2009, the annual average of patent values fluctuated up and down. However, since 2009, the annual average of patent values has been decreasing year by year. Overall, the average patent value is also declining. The significant fluctuations in patent values in different years reflect a correlation between the year and patent value.

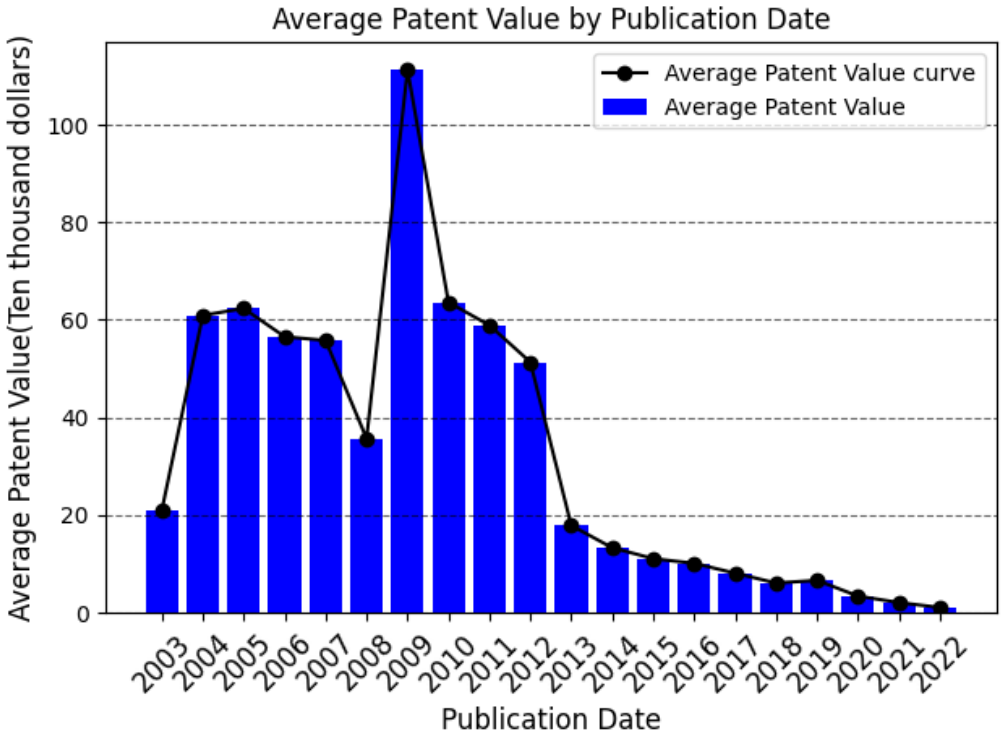


Figure 5: Average Patent Value by Publication Date

4. Industry Chain Position, Primary Technical Branch, Patent Type,IPC,Patent Validity. These indicators are typical categorical features, and Figures 4-6 display the mean distribution of patent values within these indicators.

From Figures 4 to 6, it can be observed that the range of patent values within the "Primary Technical Branch" indicator is approximately $30,000,

within "IPC" it is around $20,000, within "Patent Type" it's roughly $250,000, within "Industry Chain Position" it's about $35,000, and within "Patent Validity" it's approximately $45,000. Through ANOVA analysis, it was determined that there are no groups with the same means, indicating differences in patent values across the various indicators. Therefore, it can be concluded that these indicators are related to patent value.

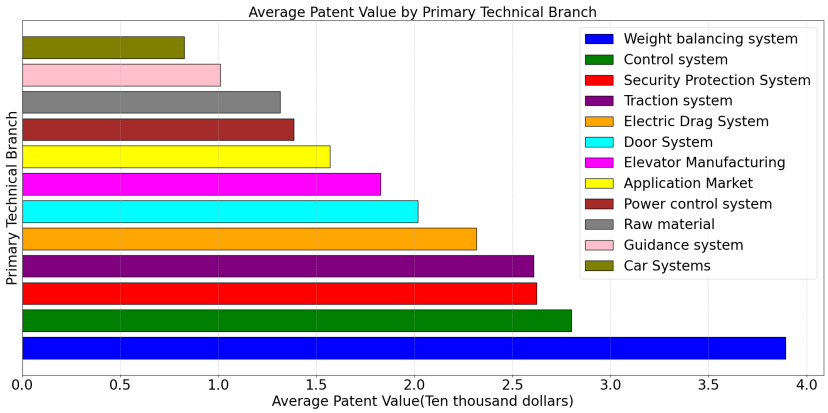


Figure 6: Average Patent Value by Primary Technical Branch

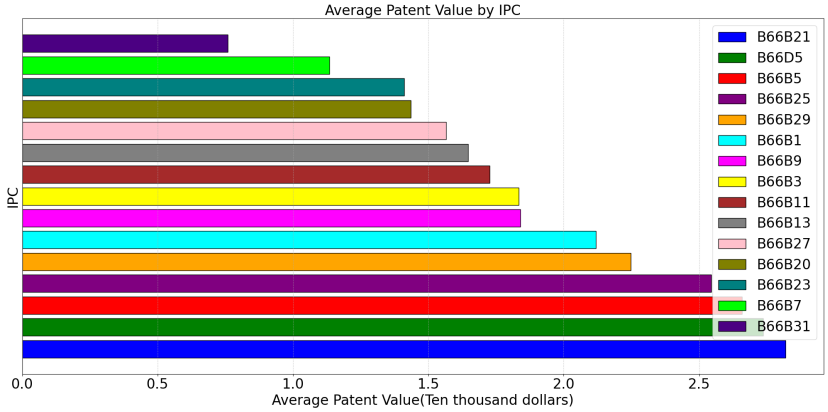


Figure 7: Average Patent Value by IPC

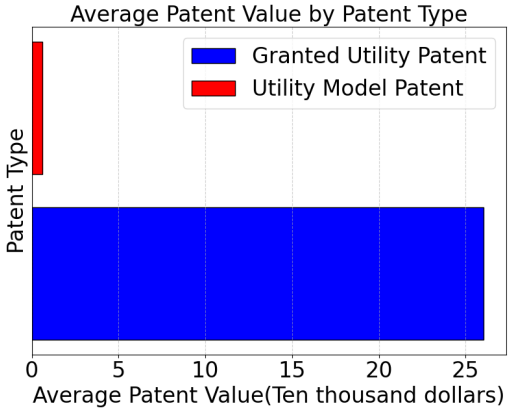


Figure 8: Average Patent Value by Patent Type

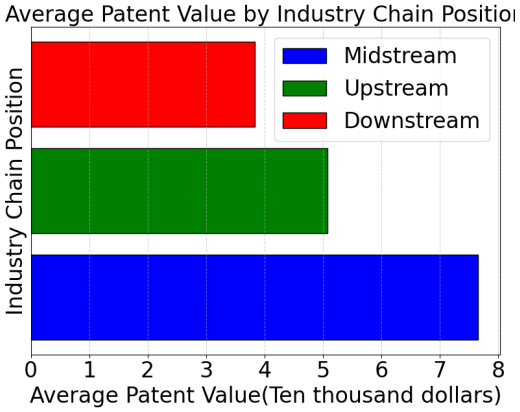


Figure 9: Average Patent Value by Industry Chain Position

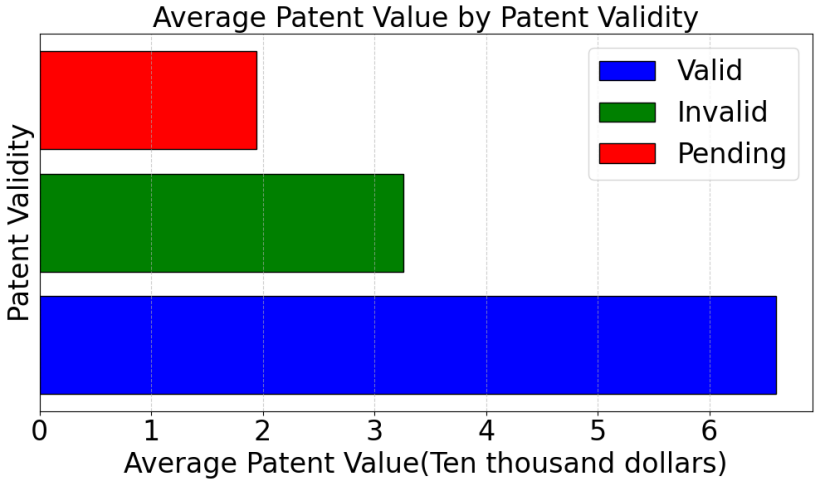


Figure 10: Average Patent Value by Validity

4.3.2. Analysis of Quantitative Features

Pearson correlation coefficient calculation formula:

$$\rho = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where x_i and y_i are individual data points in the sample; $\bar{x}$ and $\bar{y}$ are the means of x and y;the symbol $\sum$ represents the summation over all data

points.

We perform pairwise Pearson correlation analysis on quantitative features and calculate correlation coefficients programmatically. Here, we take Patent Value and Number of Claims as an example, with Patent Value as x and Number of Claims as y. First, we extract these two attributes from the dataset. Then, we calculate the means of Patent Value and Number of Claims, denoted as $\bar{x}$ and $\bar{y}$. Subsequently, we input the data points into the formula to calculate the Pearson correlation coefficient.The Pearson correlation coefficient matrix is shown in Figure 9.

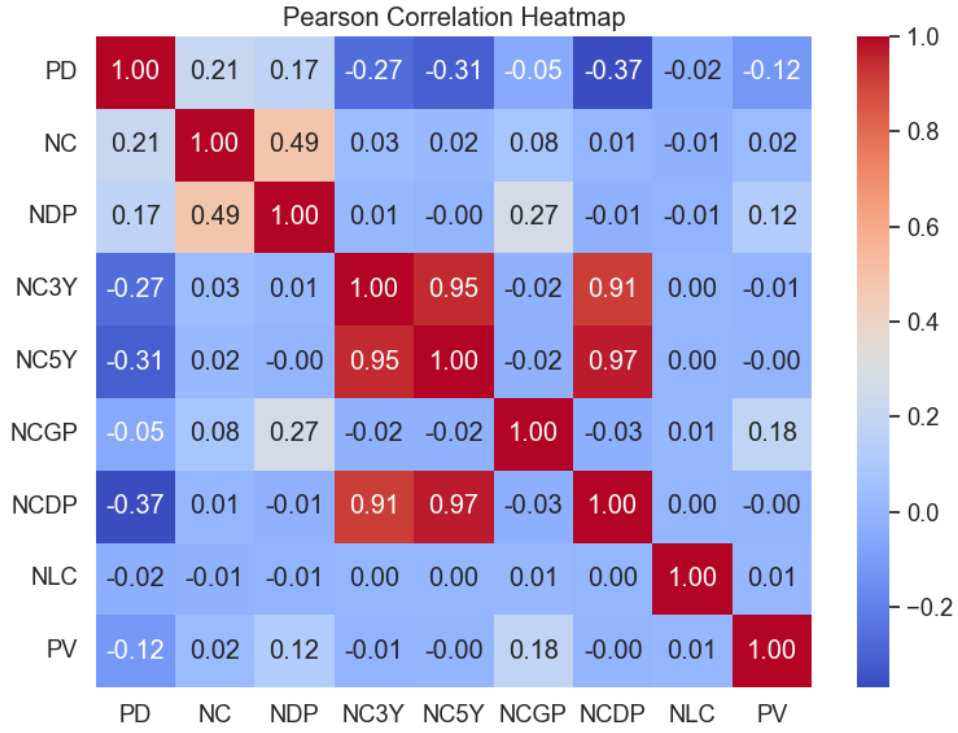


Figure 11: Pearson Correlation Coefficient Heatmap

Due to the lengthy label names, we have abbreviated them using the initials of each word. In the figure, PD (Publication Date) shows a negative correlation with PV (Patent Value), which aligns with the statistical results we obtained when treating Publication Date as a categorical feature. NDP (Number of Document Pages) exhibits a moderate positive correlation with

Patent Value. Importantly, we observed that NCGP (Number of Citing Patents) is moderately correlated with Patent Value, while NCDP (Number of Cited Patents) shows almost no correlation with Patent Value. In most studies on patent value, it is commonly believed that the more a patent is cited, the more influential and valuable it should be. However, based on our data, a patent's value tends to increase as it cites other patents more frequently.

4.4. Construction of PVCP Model

We first classified the patent data in the dataset based on patent value, adding a 'Patent Value Classification' label, with values filled as 'large' for patents valued above $10,000 and 'small' for patents valued at or below $10,000. We then performed data cleaning to handle missing values. Finally, we trained a decision tree classification model and calculated information about the number of leaf nodes and accuracy for different depths.

Table 1 illustrates the relationship between decision tree depth and the accuracy on the test set.

Table 1: the Accuracy of DTC Models with Different Tree Depths

Accuracy Max layer	Accuracy
1	96.65%
2	97.92%
3	98.39%
4	98.39%
5	98.48%

Results Analysis: From Table 1, it can be observed that the decision tree performs exceptionally well, achieving an accuracy of 99% when the tree reaches a depth of 10 layers. However, there is an unusual phenomenon where the decision tree attains a 96.65% accuracy at the first-level decision. To analyze this phenomenon, we have created process diagrams for the first three levels of the decision tree, as shown in Figure 10.

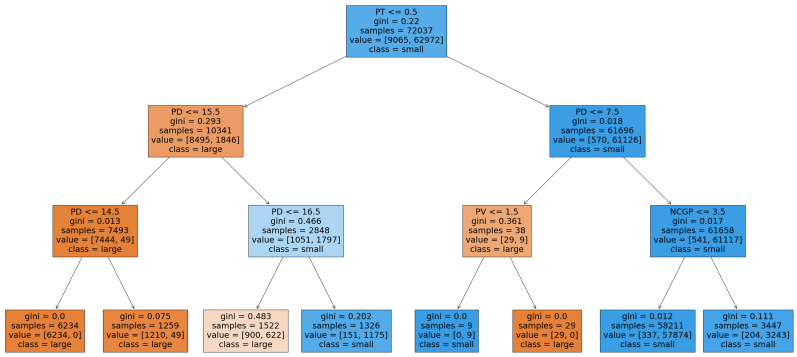


Figure 12: Decision Tree Process Diagram

From Figure 10, it was observed that the decision tree uses 'Patent Type (PT)' as a decision criterion at the first level. Therefore, we calculated the percentages of 'small value' and 'large value' under different 'Patent Type' categories, as shown in Table 2.

Table 2: the Accuracy of DTC Models with Different Tree Depths

	Granted Utility Patent	Utility Model Patent
small value	3.65%	96.35%
large value	98.44%	1.56%

5. Conclusion

References

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986. doi: https://doi.org/10.1007/BF00116251.

Yu-Jin Jeong and Byung-Un Yoon. Technology planning through technology roadmap: Application of patent citation network. *Journal of the Korea Academia-Industrial Cooperation Society*, 12(11):5227–5237, 2011. doi: https://doi.org/10.5762/KAIS.2011.12.11.5227.

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993. doi: https://doi.org/10.1145/170035.170072.

LAI Chaoan and XU Cuilu. The application of patent mining in the forecast of smart home industry. *Management Science and Engineering*, 10(1): 67–75, 2016. doi: https://doi.org/10.3968/8220.

HARRIS L J SANDERS B S, ROSSMAN J. The economic impact of patents[j]. *Trademark and Copyright Journal*, 2(2), 1958.

Manuel Trajtenberg. A penny for your quotes: patent citations and the value of innovations. *The Rand journal of economics*, pages 172–187, 1990. doi: https://doi.org/10.2307/2555502.

Mark Schankerman and Ariel Pakes. Estimates of the value of patent rights in european countries during the post-1950 period. *The economic journal*, 96(384):1052–1076, 1986. doi: https://doi.org/10.2307/2233173.

Zvi Griliches. Patent statistics as economic indicators: a survey. In *R&D and productivity: the econometric evidence*, pages 287–343. University of Chicago Press, 1998.

Peter Neuhäusler and Rainer Frietsch. Patent families as macro level patent value indicators: applying weights to account for market differences. *Scientometrics*, 96:27–49, 2013. doi: https://doi.org/10.1007/s11192-012-0870-y.

Gaston Llanes and Stefano Trento. Patent policy, patent pools, and the accumulation of claims in sequential innovation. *Economic Theory*, 50: 703–725, 2012. doi: https://doi.org/10.1007/s00199-010-0591-5.

Dietmar Harhoff, Frederic M Scherer, and Katrin Vopel. Citations, family size, opposition and the value of patent rights. *Research policy*, 32(8): 1343–1363, 2003. doi: https://doi.org/10.1016/S0048-7333(02)00124-5.

Yongtae Park and Gwangman Park. A new method for technology valuation in monetary value: procedure and application. *Technovation*, 24(5):387–394, 2004. doi: https://doi.org/10.1016/S0166-4972(02)00099-8.

Herman Wold. Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, 12 (S1):117–142, 1975. doi: https://doi.org/10.1017/S0021900200047604.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995. doi: https://doi.org/10.1007/BF00994018.

Secil Ercan and Gulgun Kayakutlu. Patent value analysis using support vector machines. *Soft computing*, 18(2):313–328, 2014. doi: https://doi.org/10.1007/s00500-013-1059-x.

George EP Box and R Daniel Meyer. An analysis for unreplicated fractional factorials. *Technometrics*, 28:11–18, 1986. doi: https://doi.org/10.1080/00401706.1986.10488093.