Journal of Informetrics 17 (2023) 101406

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

# Evaluation and identification of potential high-value patents in the field of integrated circuits using a multidimensional patent indicators pre-screening strategy and machine learning approaches

Zewen Hu[a], Xiji Zhou[a], Angela Lin[b,*]

[a] School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, People's Republic of China
[b] Information School, University of Sheffield, Sheffield S10 2TN, United Kingdom

A R T I C L E I N F O

*Keywords:*
High-value patents
Zero-cited patent
Machine learning
Integrated circuits
Patent indicator portfolio
Automatic classification
First-citation speed

A B S T R A C T

Early identification of high-value patents has strategic and technological importance to firms, institutions, and governments. This study demonstrates the usefulness of the machine learning (ML) method for automatically evaluating and identifying potential high-value patents. The study collected 31,463 patents in the integrated circuits sector using the DII platform and used them to conduct experiments using five standard ML models. A multidimensional value indicator portfolio was established to measure patents' legal, technological, competitiveness, and scientific values and construct feature vector space. The portfolio also formed a part of the pre-screening strategy providing a valid positive sample for identifying potential high-value patents. The results suggest that the multidimensional patent indicator portfolio effectively measured patent values. amongst all indicators, patent family size (legal value), first citation speed (technological value), forward citations and extended patent family size (competitiveness value), length of the patent document, non-patent reference count, and patent citation count (scientific value) play a significant informing role in identifying potential high-value patents. The proposed first-citation speed indicator proved valuable for measuring patents' technological value. The Random Forest model had the best overall performance in classifying and recognizing potential high-value patents(PHVPs) with accuracy and precision rates above 95%.

## 1. Introduction

Patents have economic and strategic importance because patented inventions' economic and technological value can impact subsequent technological development (Squicciarini et al., 2013). In this respect, patented inventions as economic resources are critical to a firm's or country's technological development. Effective patent evaluation and identification of high-value patents (HVPs) can inform decision-makers concerning investment in technology and patent applications (Fasi, 2022). Likewise, a government can formulate its science and technology policies to increase the country's competitiveness and stimulate economic growth (Chen et al., 2007). Therefore, assessing patents' value is of interest to academics and practitioners.

Varieties of methods have been proposed for patent evaluation. These methods can broadly be categorized into market-based and patent-based. The market-based methods evaluate a patent using the appropriate market value over similar or substitute inventions

* Corresponding author.
*E-mail address:* a.lin@sheffield.ac.uk (A. Lin).

https://doi.org/10.1016/j.joi.2023.101406
Received 11 April 2022; Received in revised form 19 February 2023; Accepted 6 April 2023
Available online 19 April 2023

in the market. The assumption is that the value of a patent is reflected in its observable commercial performance and, thus, in its market value (Caviggioli et al., 2020). An advantage of the market-based method is that the patent value is quantified, presented in currencies, and comparable to other patents and assets.

Evaluating patents with market-based methods can be challenging. The critique argued that assessing a patent's market or economic value requires a new approach that generates the monetary value of technology based on the structural relationship between technological and market factors (Park & Park, 2004). The critiques also argue that a patent's market value is not always observable because patent rights are not frequently traded in open markets (Kabore & Park, 2019). Marketplaces for patent transactions, such as Ocean Tomo, offer periodical patent auctions to facilitate observing a patent's private value. However, due to confidentiality, these marketplaces cannot publicize patent trades and only offer a few patent transaction data. Even when the trading price of a patent is available, the price does not always reflect the technological value of the patent (Choi et al., 2019).

The patent-based approach uses patent indicators to evaluate patent values. The approach focuses on patent quality and relies on patent-based proxies for the technological value of innovation. The premise is that better patent quality leads to a more significant impact on sequent technology development and yields higher economic benefits (Chen et al., 2007).

HVPs usually have high economic and technological value (Squicciarinim et al., 2013). Nevertheless, it is noticed that HVPs also have high values in other dimensions such as legal, competitiveness, strategic, and scientific value. Therefore, using a wide array of indicators to capture patent value in different dimensions can better reflect the overall value of a patent (Fischer & Leidinger, 2014).

Given the economic and technological importance of HVPs and patent evaluation methods, the study aimed to identify an automatic evaluation and identification method using a machine learning (ML) approach. To achieve the aim, the study addressed the following research questions.

- What value dimensions can the value of a patent be measured from?
- What methods can be used for the early identification of high-value patents?
- What strategies can be used to develop the methods?
- Which method can deliver the overall best outcome?

The study used mixed methods combining patent-based bibliometrics and machine learning (ML). The pre-screening approach based on a multidimensional value indicators portfolio was deployed to pre-evaluate HVPs from loads of patens to lay the foundation to ensure the robustness of the ML models. The research context was the integrated circuits (IC) since the field is a frontier of science and technology where patents are critical to technology development. Thus early identification of high-value patents can be beneficial.

## 2. Related work

### *2.1. Research on patent evaluation*

Extensive efforts have been made to explore indicators to evaluate patent values (Table 1). Squicciarini et al. (2013) assessed the robustness of 11 patent indicators for evaluating the economic and technological value of patented inventions. Grimaldi and Cricelli (2020) conducted a systematic literature survey and identified 37 indexes, and grouped them into five categories defined by IPScore 2.0®: Legal Status, Technical, Market conditions, Strategy, and Finance. Cricelli et al. (2021) proposed a patent value evaluation framework containing 41 key performance indicators.

The portfolio approach using multiple indicators is more effective because patents have multifaceted values (Fischer & Leiginger, 2014). Ploska et al. (2019) used eight criteria simultaneously for patent ranking and evaluation. They found that only some essential patents can be identified with configurations with less than five criteria. Conversely, the configurations with more than or equal to five criteria can find almost all the important patents. Based on the results, Ploska et al. argued that the number of criteria used for evaluation and ranking patents relates to the accuracy of patent identification. Hsieh (2013) developed a technology portfolio planning model which combined indicators from two dimensions (risks and benefits) to evaluate patent value. Grimaldi et al. (2015) proposed a three-dimensional patent indicator portfolio including patent strategy, business value, and technology to assess patent value. Van Raan (2017) suggested that patent quality assessment can be best based on a combination of different economic value indicators of patents, and the combination depends on the dimension(s) of the patent value it wants to measure. Wu et al. (2021) argued that the weighting given to indicators and value dimensions relies on expert judgments, and such a decision is subjective. They established an index system to evaluate patent quality, consisting of indicators from technical, legal, and market condition value dimensions.

### *2.2. Patent evaluation methods*

Patent citation analysis is a commonly used and considered effective method and can help to identify the scientific intensity of a technology (Callaert et al., 2006; Xu et al., 2021; Liu et al., 2021), technological trajectories (Huang et al., 2020; Lai et al., 2021), technological breakthroughs in a field (Albert et al., 1991), radical innovations (Arts et al., 2012), the technological improvement rate of a field (Benson & Magee, 2015), emerging technologies (Chi & Wang, 2022), promising technologies (Noh & Lee, 2022), and potential technological opportunities (Lee et al., 2018; Seo, 2022).

ML emerged as a patent value evaluation and identification tool to process, analyse, and evaluate a large volume of patents. The models can automate the patent evaluation process to replace total or partial expert evaluation, done manually (Lee et al., 2018; Kwon & Geum, 2020; Trappey et al., 2019). Hido et al. (2012) used ML and text-mining methods to compute a score to indicate how likely

**Table 1**
Examples of the research on patent evaluation.

| No. of dimensions | Value dimension | Patent indicators | References |
|---|---|---|---|
| 1 | Technical | Claims, classification, patent family, forward citations | Liu, Qiao, & Liu, 2021 |
| | Technical | Forward and backward citations | Lai et al., 2021 |
| | Economic | The market size of the patent family | Kabore & Park, 2019 |
| 2 | Technology and Market value | *Technology:* proprietary position, level of technology, & life of technology, degree of standardization, type of technology, contribution ratio, scope of application, and degree of completeness. *Market:* profit-generating, cost-saving, amount of income, duration of income, and risk of income. | Park & Park, 2004 |
| | Technological competitiveness and innovation force | *Technological competitiveness*: TR Patent Scorecard. *Innovation force*: essential patent index and essential technology strength. | Chen et al., 2007 |
| | Internal and external indicators | *Internal indicators*: technological scope, priority range, geographical scope, cooperation degree, completeness. *External*: IPC perspective, backward citation perspective | Choi et al., 2019 |
| 3 | Technological quality, economic relevance, and patent scope | *Technological quality:* number of forward citations. *Economic relevance*: family size. *Patent scope*: number of IPC classes | Fischer & Leidinger, 2014 |
| | Technical value, legal value, and Market value | *Technical:* technical quality, enforcement, & applications range. *Legal:* scope of rights protection, stability of rights. *Market:* market application, patent trading, future market expectation. | Wu et al., 2021 |
| | Economic, technical, and legal value | *Economic*: market application, enterprise patents, and sales ratio. *Technical*: number of inventors, number of citations, number of classifications, be cited, number of claims. *Legal*: number of siblings, manual pages, survival period, license status. | Huang et al., 2022 |
| 4 | Technology features, technology transfer specifications, market status, and legal strategies | *Technology features* technology maturity, technology innovation and technology background. *Technology transfer*: R&D capability of the technology transferors and transferees. *Market status*: the market scale and market segmentation. *Legal strategies*: property ownership and authorization specifications of technology. | Hou & Lin, 2006 |
| 5 | Technological, radical, economic, social and strategic value | *Technological:* the degree to which a patent contributes to the subsequent development of the technology. *Radical and incremental*: the technological value of a patent is also measured by if the patented invention is a start of new technology (radical) or an improved version of the existing technology (incremental). *Economic*: economic or commercial benefits of the patented invention brought to the firm. *Social*: how a community or society benefit from technological development. *Strategic*: how a patented invention protects a firm from its competitors and the market. | Frietsch et al., 2010 |

an application will be approved to save patent evaluators' time. The authors suggested that their proposed method outperformed the traditional method, which considered only the structured properties of the documents. Lin et al. (2018) proposed a Deep Learning-based Patent Quality Value model to evaluate new patents' quality. They argued that the conventional method, such as patent citation analysis, might overlook crucial information (e.g., text materials). Trappey et al. (2019) also applied a deep learning analytical method for automating the estimation of the patent value in the IoT field. They first used principal component analysis to identify significant patent value indicators from the given dataset. A total of eleven value indicators were selected and then used in a Deep Neural Network (DNN) for value prediction. The results showed that the DNN model performed better in accuracy than the traditional Back-Propagation Neural Network technology.

### *2.3. Patent value indicators*

This study viewed HVPs as characterized by high legal, technological, competitiveness, and scientific values. Therefore, it used a portfolio approach containing indicators from these four dimensions to evaluate and identify PHVPs. The study selected the most commonly used and whose information is easily accessible value indicators to establish the portfolio (Table 2).

#### *2.3.1. Legal value*

*2.3.1.1. Patent claims.* Claims are the basic unit for measuring the inventiveness of patents. According to the United State Patent and Trademark Office (USPTO) ,[1] the claims must satisfy "the subject matter that the inventor or inventors regard as the invention" and "defines the scope of the protection of the patent." There are two types of patent claims: independent and dependant (Moehrle & Frischkorn, 2021). The former is a broader definition of the scope of the patent, and the latter depends on the former, which usually is a specific description of the elements on which the former depends (Yang & Soo, 2012). A typical patent consists of several claims, each representing a distinctive invention. A patent's claim entails the legal definition of an invention and gives exclusive rights to the invention protected by law. The number of claims in a patent indicates the scope and breadth of legal protection. Hence, the larger the number of claims, the broader the scope of legal protection and the more resilient and longer-lived the patent is (Trappey et al., 2012). Longer-lived patents are generally more valuable (Choi et al., 2019; Kabore & Park, 2019).

[1] USPTO, Nonprovisional (Utility) Patent Application Filing Guide, 2009. https://www.uspto.gov/patents-getting-started/patent-basics/types-patent-applications/nonprovisional-utility-patent#heading-18. (Accessed 1 August 2017).

**Table 2**
Definition of patent value indicators.

| Value dimension | Indicator name | Definition | Perceived value |
|---|---|---|---|
| Legal | Claims | Legal breadth, the subject of legal protection for patent holders | Positive for a higher count |
| | Patent family size | The number of jurisdictions where patent protection has been sought | Positive for a broader coverage |
| Technological | First-citation speed | The time lag between patent publication and the first forward citation received | Positive for a shorter time |
| | Classification scope (IPC categories) | Number of IPC classification | Positive for a higher coverage |
| Competitiveness | Forward citations | Citations received from other patents under application | Positive for a higher citation count |
| | Extended patent family size | Subsequent inventions built on the patent-protected invention | Positive for larger size |
| | - Inventor team size<br>- Number of patent holders | - Number of inventors<br>- Number of patent holders and patent owners | - Positive for larger size<br>- Positive for a higher count |
| Scientific | Length of patent instruction document | Page counts of patent instruction | Positive for a higher count |
| | Patent reference count | Number of past patents cited by patent application and given by examiners | Positive for a higher citation count |
| | Non-patents reference count | Citations of non-patent references | Positive for a higher citation count |

*2.3.1.2. Family size.* Patent family size refers to "the total number of jurisdictions in which patent protection is sought for the same invention" (Harhoff et al., 2003). Patent families indicate a company's innovation strategy and patent portfolio and confidence in receiving a return on investing in the patent (Wu et al., 2015). The patent family size also indicates the invention's importance since the patent claims are recognized by the countries which granted the exclusive rights. Thus, the greater the number of countries where a patent is granted, the larger the size of the patent family and the stronger the legal protection of patent rights and infringement litigation (Kabore & Park, 2019).

### *2.3.2. Technological value*

*2.3.2.1. First-citation speed.* There is a distinction between short- and long-term forward citation concerning a patent's technological value and importance. Short-term forward citations refer to citations received within the first five years of the patent being granted, whereas long-term citations are those received beyond this period (Lanjouw & Schankerman, 2001). Although short-term patent citation can effectively solve the right truncation problem of patent citation (Fisch et al., 2017), the utility of the first citation of patents is different from that of subsequent citations. Studies have shown that the probability of patents being cited reaches its peak at its "young" patent age (Mehta et al., 2010). In addition, important patents often receive more citations quickly, accompanied by a faster citation rate (Doonan et al., 2019).

The time lag between the granting of a patent and its first forward citation indicates the technological value of the patent (Lee & Sohn, 2017). In general, the quicker a patent receives its first citation, the more forward citations it will receive subsequently (Gay et al., 2005) and the more important it is (Mehta et al., 2010). We term this time lag *first-citation speed* and argue that the faster the first-citation speed, the greater the technological value of the patent.

*2.3.2.2. Classification scope.* Classification scope, such as number of IPC categories, measures the technological breadth. The classification scope indicates the diversity of knowledge cutting across different technological fields and diversified technical features of an invention. An increase in IPC categories means that patented inventions span a single technical field and have diversified technical features (Lee & Sohn, 2017; Park & Yoon, 2017). Thus, the number of IPC classes assigned to a patent indicates the importance of the patent. The IPC technology category can measure the proximity of patented technology in the technical field, and the IPC subcategories (4 digits of the classification system) are often the criteria for technical evaluation.

### *2.3.3. Competitiveness value*

*2.3.3.1. Forward citations.* Forward citations are the citations that a given patent receives and are an essential proxy of patent value (Arts et al., 2012; Benson & Magee, 2015). Squicciarini et al. (2013) argued that the number of forward citations "mirrors the technological importance of the patent for the development of subsequent technologies, and also reflects, to a certain extent, the economic value of inventions" (p. 35). Forward citation count evaluates a patent's innovation, economic, and market value (Zhang et al., 2012; Ruan et al., 2020; Ruan et al., 2021). The higher the technological quality of a patent, the more inventions will build on it, thereby, the greater the value of the patent's exclusive right. Thus, the more forward citations a patent receives, the higher its contribution to subsequent inventions and the higher its value.

*2.3.3.2. Extended patent family size.* The size of the extended patent family indicates the number of subsequent inventions built on a particular patented technology. The larger the size, the more comprehensive the coverage of the patented technology and the more prominent the strength of rights protection and competitive advantages (Frietsch & Schmoch, 2010). The size of the extended patent family also shows a company's ability to span and integrate industrial, market, and regulatory factors to cover the temporal, geographical, technological, and product dimensions to develop its patent portfolio. The extended patent family size of technology ultimately points to its competitiveness.

*2.3.3.3. Inventor team size and number of patent holders.* Inventor team size consistently correlates with patent value and technological usefulness (van Zeekbroeck & van Pottelsberghe de la Potterie, 2011). The indicator reflects the knowledge foundation of an invention. Therefore, the larger the inventor team size, the more solid the patent research and development knowledge base, and the greater possibility that the invention is highly competitive (Munari & Oriani, 2011). Kiehne and Krill (2017) found that the number of inventors per invention positively correlated with patent value. Breitzman and Thomas (2017) found that inventor team size is a good indicator of predicting the future citation impact of patents.

The number of patent holders also reflects the resource allocations to an invention. In general, a more significant number of patent holders indicates more investment, maintenance, and operation status, which is conducive to the continuous realization of the value of the patent (Kim et al., 2021)

### *2.3.4. Scientific value*

*2.3.4.1. Length of patent instruction document.* Patent instruction contains technical descriptions, specifications, claims scope, and other information. The length of the patent instruction is positively correlated with patent value since the document presents a patent's technical scope, details, and performance (Reitzig, 2004). Trappey et al. (2012) used the lengths of detailed specifications to measure the value of a patent. Barney (2010) argued that a more extended specification provides better support for patent claims, strengthening the patents against possible validity attacks.

*2.3.4.2. Non-patent references.* A patent may be partially or entirely based on new scientific knowledge, and scientific papers are the source of new knowledge and evidence to support inventions built on that knowledge. As early as 1981, Carpenter's team proposed that the citation of scientific articles is an important indicator to reflect the novelty of patents because patents that directly rely on scientific literature can better reflect their close scientific connection than those that are only based on existing technology. Callaert (2006) and Nagaoka (2007) pointed out that more scientific references indicate that patented technologies are more closely related to scientific activities and are positively related to patent value.

*2.3.4.3. Patent references.* Backward citations are citations of existing and relevant technologies upon which the patented technologies draw. Backward citations evaluate the novelty of a patented invention. Moreover, backward citations can assess the degree of knowledge linkage within and between the technical fields and is an intuitive reflection of the accumulation of patent innovation knowledge (Lee, 2018). Backward citations can also evaluate an invention's patentability and define the legitimacy of the claims in the application (Squicciarini et al., 2013). The number of backward citations implies the number of sources upon which a given invention is built. Patents with a more significant number of backward citations tend to have greater market values because the citation numbers indicate how solid the technological base of the invention is. Bessen (2008) pointed out that if a patent cites many other patents in the technical category, it will have a high originality index.

## 3. Methodology

Fig. 1 presents a methodology for evaluating and identifying PHVPs. The column on the left shows the theoretical foundations informing the value indicators selection (Section 2) and machine learning model constructions (Section 3.4). The column on the right shows the core activities involved in the process. The remainder of the section discusses each of these core activities in turn.

### *3.1. Constructing a patent database*

State of the US Semiconductor Industry (2021) described IC as the brain of the new generation of information technology. The industrial chain of IC is involved in almost all modern technologies, including 5 G, artificial intelligence (AI), quantum computing, the Internet of Things (IoT), etc. Thus, the IC field, which has many HVPs, provides an ideal context for the study to test its proposed method.

Fig. 2 depicts a systematic and incremental search approach adopted to ensure systematic patent data collection (Huang et al., 2011). A comprehensive search across Encyclopaedia Britannica and Wikipedia using the search term 'integrated circuit' was performed to identify the terminologies associated with IC and their definitions. For example, "integrated circuit (also integrated microcircuit), a microminiaturized electronic device…", "integrated circuit– The formal name of the chip.", "integrated circuit (IC, or 'chip') A microelectronic semiconductor device consisting of…". Given the lag in updating encyclopaedia entries, relevant websites and forums, such as JEDEC Solid State Technology Association and Elprocus Engineering Electronics Project, were also searched.

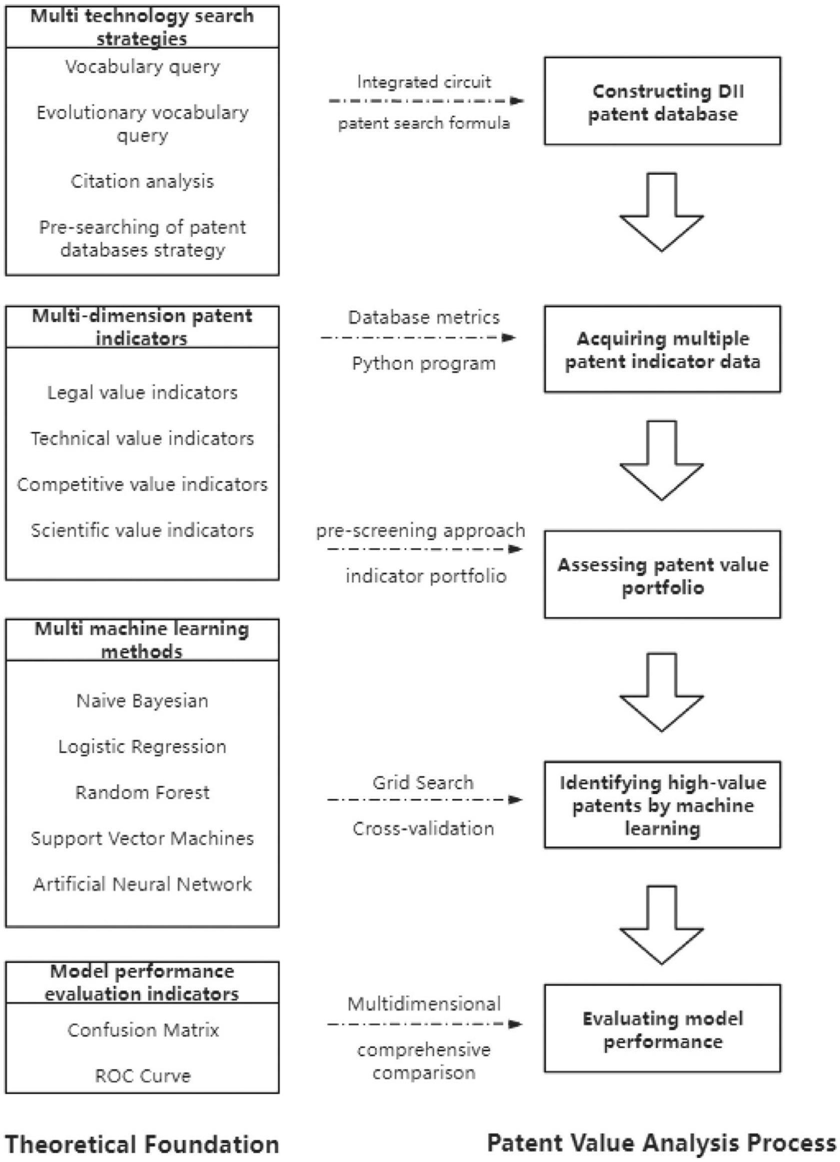


**Fig. 1.** High-value patent evaluation and identification methodology process flow.

Then, Derwent Innovations Index (DII) and Incopat databases were searched. Each search generated a set of IC-related terminologies, and all terminologies were consolidated and turned into the advanced search query.[2]

The search for patents published in the IC field in 2015 was conducted on July 28, 2021, and the search was limited to the patent titles and abstracts in DII to eliminate irrelevant records. A total of 32,302 patents were retrieved.

### 3.2. Acquiring multiple patent indicator data

The steps taken to extract and prepare multiple patent indicator data exported from the DII were:

*Simple counting.* Simple counting of 6 indicators: number of patent claims, extended patent family size, number of inventors, number of patent holders, instruction pages, and frequencies of backward citations.

[2] The advance search query used on DII is: TS= (integrated circuit OR IC OR microcircuit OR microchip OR chipset OR semiconductor chip OR IC chip OR integrated circuit chip OR integrated circuit chip OR unicircuit OR molectron OR integrated semiconductor) AND Publish Date = [2015-01-01 to 2015-12-31]

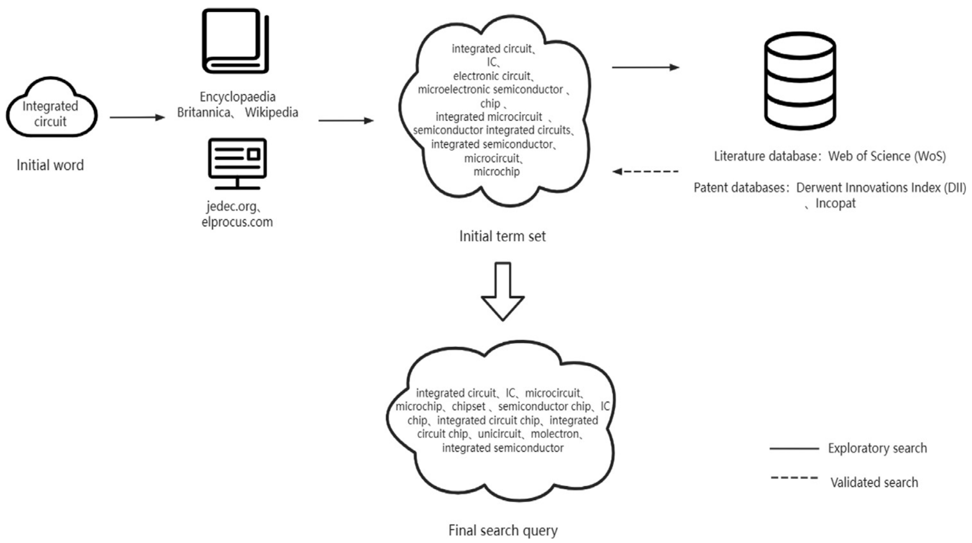


**Fig. 2.** Incremental approach to search.

*De-duplicating patent data*. The statistical unit in the DII database is per patent in a patent family, and a patent family is the aggregation of patent data of the entire family. Hence, data de-duplication is necessary. Four areas needed to be de-duplicated, including IPC classification, authorized countries, citing patents and citing literature.

*Making the statistics*. Python program collected information on times that patent families received citations. This was used to calculate the time interval between the publication year of the priority patent of a patent family and the year that the patent family was cited for the first time.

*Descriptive statistics*. Descriptive statistics show the overview of the data, including the distribution range and degree of dispersion. The statistics show abnormal patent data (e.g., patent instructions with zero page) and the deviation and stability of the patent data values to ensure the quality of index data. Excluding 839 abnormal patent data, the final 31,463 patent data are used for evaluating and identifying patent value.

### 3.3. Evaluating the patent value using a pre-screening approach

This study pre-evaluated patent value using the multidimensional value indicators portfolio presented in Table 2. The multi-dimensional value indicators portfolio can reflect the multifaceted characteristics of high-value patents. The existing patent value evaluation standards that focus on specific thresholds or classes of a single index cannot fully describe high-value patents' features. The pre-screening step lays the foundation for the subsequent automatic identification and classification of high-value patents using machine learning.

### 3.4. Machine learning identification methods for potential high-value patents

Given that different ML models have strengths and weaknesses in patent data processing and HVPs identification, this study selected Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine, and Artificial Neural Network for comparison.

#### 3.4.1. Naive bayes

The Naive Bayes (NB) model identifies high-value patents through the maximum-likelihood training to obtain the sample's final classification using formula (1).

$$P(\theta|D) = \frac{P_0(\theta) * P(D|\theta)}{P(D)} \quad (1)$$

Above

Given the model parameters $\theta$ and the data set D

$P_0(\theta)$ is the prior probability distribution

$P(D)$ is the edge likelihood function which is constant

$P(D|\theta)$ is the likelihood function (or conditional probability function), which also conforms to the principle of the prior probability distribution

$P(\theta|D)$ is the maximum value of the posterior probability distribution taken as the category X belongs to given the feature vector$x_1, x_2, x_3, \cdots, x_m$ with $m$ number of features.

### 3.4.2. Logistic regression

Logistic regression (LR) uses a logistic function to map the output value to (0, 1) and determine the classification of the sample by setting a threshold. Formula (2) combines the standard logistic Sigmoid function and regression model and assumes that the data obeys the Bernoulli (0–1) distribution to model the probability of a given sample.

$$\begin{cases} P(Y=1|x) = \frac{1}{1+e^{-w^T x+b}} \\ P(Y=0|x) = 1 - P(Y=1|x) \end{cases} \tag{2}$$

Above

*$x$ is the input vector of the sample set $Xw$ is the weighted vector, and b is the bias term, which is often solved by maximizing the likelihood function or minimizing the loss function$P(Y=1|x)$ and $P(Y=0|x)$ refer to the probability that the category event Y label is 1 and 0 under the given feature vector of the sample $x$, respectively. The instance x will be assigned to the larger probability value category.*

### 3.4.3. Random forest

Random Forest (RF) performs a classification task by constructing multiple CART decision trees to carry out feature selection based on the Gini coefficient. The smaller the Gini coefficient, the higher the feature purity of each sub-node and the smaller the data uncertainty.

$$Gini(p) = \sum_{K-1}^{K} p_K \left(1 - p_K\right) \tag{3}$$

$K$ is the number of sample categories

*$p_K$ is the probability of the sample being correctly classified into the $K_{th}$ category$1 - p_K$ is the probability of the sample being misclassifiedGini$(p)$ is the sum of all classification error rates in the sample set.*

### 3.4.4. Support vector machine

There are two types of Support Vector Machine (SVM): linear and nonlinear. Linear SVM is applied when data is perfectly linearly separable, and nonlinear SVM is applied when data is not linearly separable.

$$\begin{aligned} &\max_{w,b} \frac{2}{\|w\|} \\ &s.t. y_i\left(w^T \cdot x_i + b\right) \geq 1, i = 1, 2, \cdots, m \end{aligned} \tag{4}$$

Formula (4) is linear SVM where $w$ and $b$ are the separation hyperplane parameters, and the maximum separation hyperplane is $max\frac{1}{\|w\|}$. When the original sample space is linearly inseparable, SVM maps the sample to the high-dimensional feature space by introducing a kernel function, and obtains formula (5), $\phi(x_i)$ represents the feature vector after $x$ mapping.

$$\begin{aligned} &\max_{w,b} \frac{1}{2} \| w\|^2 \\ &s.t. y_i\left(w^T \cdot \phi\left(x_i\right) + b\right) \geq 1, i = 1, 2, \cdots, m \end{aligned} \tag{5}$$

### 3.4.5. Artificial neural network (ANN)

Formula (6) describes a three-layer neural network structure with a single hidden layer of N neurons. Each successive three-layer structure in a neural network can be viewed as the following form.

$$\begin{aligned} &element\ y = f\left(wx + b_0\right)\theta + b_1 \\ &w \in R^{N\times M},\ b_0 \in R^{1\times N}, \theta \in R^{L\times N}, b_1 \in R^{1\times L} \end{aligned} \tag{6}$$

Above

*$x$ ($x \in R^{1\times M}$) is a vector consisting of M input features$y$ ($y \in R^{1\times L}$) is a vector containing L output variables$w$ and $\theta$ are the weight matrices from the input layer to the hidden layer and the hidden layer to the output layer, respectively$b_0$ and $b_1$ are the constant vectors of bias terms of each neuron in the hidden layer and each variable in the output layer, respectively$f(x)$ is the activation function, often in the form of a step function, sigmoid function, etc.*

### 3.4.6. Model evaluation

To improve the training and generalization capability of the classification models, parameter values in each model were traversed by methods such as network search. The K-fold cross-validation method was used to divide the dataset. In addition, the Confusion Matrix was used to measure the performance of the model, indicators including Accuracy, Precision, and Recall. The F1 score (in Table 3) was introduced to the accuracy rate and the recall rate because these two indicators affect each other. Finally, the classifier's performance is visualized through the area under the ROC curve.

## 4. Implementation

This section describes and discusses the process of implementing machine learning models. The process is presented in Fig. 3.

**Table 3**
Model performance evaluation indicators.

| Indicator | Calculation formula |
|---|---|
| Accuracy | $\frac{TP+TN}{TP+TN+FP+TN}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| Recall | $\frac{TP}{TP+FN}$ |
| F1 score | $2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$ |
| AUC | Area under ROC curve |

*Notes:* The indicator data are all from the confusion matrix: TP is positive samples predicted to be positive by the model; FP is negative samples predicted to be positive by the model; FN is positive samples predicted to be negative by the model; TN is negative samples predicted to be negative by the model.

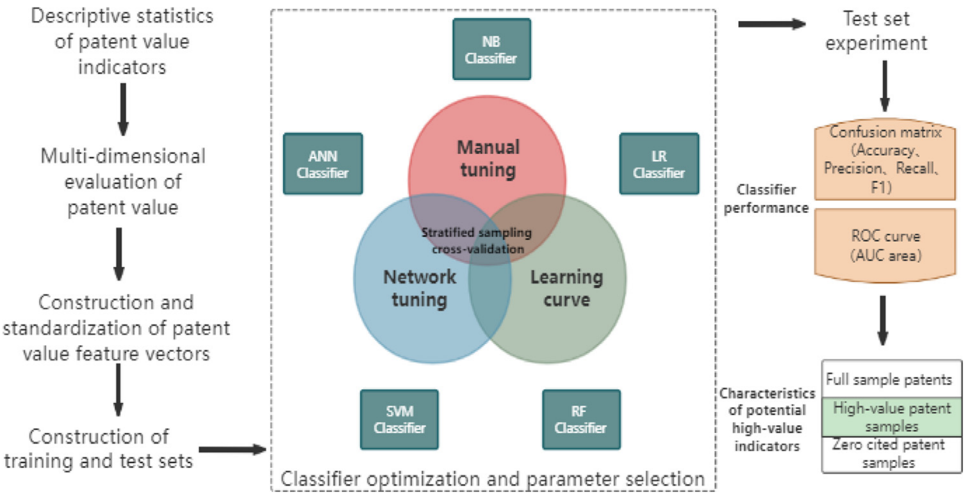


**Fig. 3.** The process of implementing machine learning models to evaluate and identify potential high-value patents.

**Table 4**
Evaluating and pre-screening of high-value patents ($n$ = 31,463).

| Value dimension | Patent indicator | Indicator mark | Max | Min | Mean | Std. | Assessment number | Numerical performance |
|---|---|---|---|---|---|---|---|---|
| Legal | Claims | $X_1$ | 51 | 1 | 3.875 | 2.090 | 1246 | [8, 51] |
| | Patent family size | $X_2$ | 144 | 1 | 16.603 | 40.925 | 1474 | [133, 144] |
| Technological | First-citation speed | $X_3$ | 28 | −4 | 14.947 | 12.590 | 1349 | [−4, 0] |
| | IPC categories | $X_4$ | 13 | 1 | 1.703 | 1.165 | 939 | [5, 13] |
| Competitiveness | Forward citations | $X_5$ | 312 | 0 | 1.748 | 6.104 | 1522 | [7, 312] |
| | Extended patent family size | $X_6$ | 59 | 1 | 2.519 | 3.025 | 1330 | [10, 59] |
| | Inventor team size | $X_7$ | 43 | 1 | 3.170 | 2.726 | 1389 | [9, 43] |
| | Number of patent holders | $X_8$ | 31 | 1 | 1.998 | 1.804 | 1053 | [7, 31] |
| Scientific | Length of patent instruction document | $X_9$ | 584 | 1 | 16.97 | 21.025 | 1393 | [48, 584] |
| | Patent reference count | $X_{10}$ | 997 | 0 | 11.906 | 42.894 | 1570 | [38, 997] |
| | Non-patents reference count | $X_{11}$ | 485 | 0 | 2.398 | 14.483 | 1451 | [9, 485] |

### *4.1. Multidimensional evaluation and pre-screening of high-value patents*

When being measured using the first-citation speed indicator, the patents were divided into two categories based on whether they received the first citation in the year of publication. For those that received their first citation in the year of publication, their first-citation speed is 0; and for those that received no citation during the given citation window, the maximum value is +1. Table 4 presents the descriptive statistics of the 11 patent indicators, including the max-min values, the mean, the standard deviation, the assessment number, and the data dispersion range.

The study pre-screened the top 5% of data in each value indicator to form its high-value patent sample. A total of 8946 high-value patents out of 31,463 patent data were pre-screened after removing data duplication. The first-citation speed is an indicator with a negative value. Thus, the smaller the value, the more it reflects the competitiveness of a patent. In this respect, pre-screening the top 5% of patents for the first-citation speed was in reverse order. The penultimate column (Assessment number) and the last column (Numerical performance) present the number and numerical performance of high-value patent samples selected for each indicator.

**Table 5**
Patent value feature vector and labelling of the training set in the field of integrated circuits.

| | **Value characteristics** | | | | | | | | | | | **mark** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **NO.** | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $Y_i^+/Y_i^-$ |
| $P_1$ | 1 | 136 | 0 | 3 | 3 | 8 | 7 | 9 | 178 | 18 | 212 | $Y_1^+$ |
| $P_2$ | 3 | 1 | 0 | 3 | 3 | 8 | 6 | 5 | 13 | 18 | 2 | $Y_2^+$ |
| $P_3$ | 5 | 1 | 2 | 2 | 3 | 5 | 6 | 6 | 16 | 18 | 0 | $Y_3^-$ |
| $P_4$ | 3 | 117 | 2 | 3 | 21 | 12 | 5 | 4 | 46 | 33 | 4 | $Y_4^+$ |
| $P_5$ | 2 | 41 | 3 | 4 | 5 | 5 | 2 | 5 | 87 | 33 | 7 | $Y_5^-$ |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| $P_{1000}$ | 5 | 132 | 0 | 3 | 2 | 9 | 17 | 3 | 23 | 33 | 4 | $Y_{1000}^+$ |
| $P_{1001}$ | 4 | 1 | 28 | 1 | 0 | 1 | 1 | 1 | 11 | 0 | 0 | $Y_{1001}^-$ |
| $P_{1002}$ | 3 | 1 | 3 | 2 | 2 | 1 | 1 | 1 | 15 | 7 | 0 | $Y_{1002}^-$ |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| $P_{18037}$ | 3 | 1 | 28 | 3 | 0 | 1 | 1 | 1 | 6 | 5 | 0 | $,Y_{,180,37}^-$ |

**Table 6**
Five machine learning models and optimized parameters.

| Classifier | Optimization method | | | | optimized | |
|---|---|---|---|---|---|---|
| | Cross-validation | Network tuning | Manual tuning | learning curve | parameter | |
| NB Classifier | √ | | | √ | | |
| LR | √ | √ | | | Penalty | Loss function |
| Classifier | | | | | L2 regularisation with C = 10 | lbfgs + single loop multinomial |
| RF | | √ | | √ | Number of single decision trees | Tree depth |
| Classifier | | | | | n_estimators = 10 | max_depth = 5 |
| SVM | | √ | | | Kernel function | Penalty |
| Classifier | | | | | rbf Gaussian kernel function with gamma=3 | L2 regularisation with C = 10 |
| ANN | √ | √ | √ | | Hidden layer | Activation function |
| Classifier | | | | | Single hidden layer of 100 nodes | sigmoid function + softmax optimizer |

The value span of some indicators and the standard deviation of the total sample is notable, indicating uneven value distribution. The uneven distribution made distinguishing patent values at different levels easier.

### 4.2. Construction of training and test sets for machine learning recognition

The high-value patents shown in Table 4 formed the positive target vector $Y_i^+$ in the machine learning training set, covering 7134 high-value patents. The distribution of high-value and other patents (i.e., the patents other than the high-value patents) in the entire dataset is 28% and 72%, respectively; therefore, using pure random sampling can cause serious deviations in the experimental results. The study used a stratified sampling method to extract the same proportion of the patents from other patents to ensure uniformly distributed training and test sets. The stratified sample was marked as the negative target vector $Y_i^-$ of the training set, covering 18,037 patents.

Table 5 shows the patent value feature vectors and markers of the HVPs in the training set, which covers: the number $P_i$ of each patent in the entire data set, the value feature indicator $X_i$; the identification target variables $Y_i^+$ and $Y_i^-$; and the feature index value $V_{ij}(1 \leq j \leq n)$. Here, $P_1$ is the first patent, $x_1$ is the first feature index of the patent, and $v_{11}$ is the first feature value of the first patent. The feature vector space of patent $P_i$ is expressed as $P_i = (x_1, v_{i1}; x_2, v_{i2}; \cdots; x_j, v_{ij}; \cdots; x_n, v_{in})$; $(v_{i1}, v_{i2}, \cdots, v_{in})$ is the corresponding feature indicator value; and $Y_i^+$are the target variables to be identified. To avoid the impact of large orders of magnitude differences on the accuracy of the model, it is necessary to standardize the data of each indicator and use the boundary value information of different sample attributes to scale the attributes to the [0, 1] interval. Thereby forming a standardized value feature vector space of integrated circuit patents.

### 4.3. Model training and optimization

The construction of five high-value patented machine learning classifiers was completed on the Jupyter platform using the Python Sklearn library. This process combined manual parameter tuning, network parameter tuning, and learning curve optimization of parameters to seek the optimal performance of the model test set. Table 6 summarises the specific optimization methods and optimal parameters of each model.

**Table 7**
Performance indicators of five machine learning model classifiers in identifying potential high-value patents.

| Classifier | Accuracy | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| NB Classifier | 0.848 | 0.754 | 0.694 | 0.723 | 0.802 |
| LR Classifier | 0.861 | 0.829 | 0.651 | 0.730 | 0.798 |
| RF Classifier | 0.953 | 0.978 | 0.858 | 0.914 | 0.925 |
| SVM Classifier | 0.942 | 0.924 | 0.871 | 0.897 | 0.921 |
| ANN Classifier | 0.942 | 0.942 | 0.749 | 0.893 | 0.914 |

**Table 8**
Comparison between the predicted characteristics of high-value patents and the averages of the two other patent samples.

| | High-value patents (1812) | All patent samples (31,463) | | | Zero-cited patent sample (15,048) | | |
|---|---|---|---|---|---|---|---|
| Patent value indicator | Mean | Mean | Mean difference | Mean multiple | Mean | Mean difference | Mean multiple |
| Claims | 4.390 | 3.875 | 0.515 | 1.133 | 3.886 | 0.504 | 1.130 |
| Patent family size | 44.497 | 16.603 | 27.894 | 1.595 | 10.010 | 34.487 | 4.445 |
| First-citation speed | 9.862 | 14.947 | −5.085 | 0.660 | | | |
| IPC classifications | 2.232 | 1.703 | 0.529 | 1.311 | 1.559 | 0.673 | 1.432 |
| Forward citations | 3.455 | 1.748 | 1.707 | 1.977 | | | |
| Extended patent family size | 4.688 | 2.519 | 2.169 | 1.861 | 1.843 | 2.845 | 2.544 |
| Inventor team size | 4.815 | 3.170 | 1.645 | 1.519 | 2.755 | 2.060 | 1.748 |
| Number of patent holders | 3.175 | 1.998 | 1.177 | 1.589 | 1.680 | 1.495 | 1.890 |
| Length of patentinstruction document | 27.997 | 16.97 | 11.027 | 1.650 | 14.487 | 13.510 | 1.933 |
| Patent reference count | 30.267 | 11.906 | 18.361 | 2.542 | 7.673 | 22.594 | 3.945 |
| Non-patent reference count | 7.164 | 2.398 | 4.766 | 2.987 | 1.156 | 6.008 | 6.197 |

## 5. Results

### *5.1. Prediction of potential high-value based on the pre-screening multidimensional patent indicators*

Table 7 summarises the performance of the five classifiers based on the pre-screening multidimensional patent indicators in accuracy, precision, recall, F1 score, and AUC.

NB and LR models were less effective in identifying potential high-value patents. RF, SVM, and ANN performed well in recognition. Their performances in all areas except for recall were poor compared to other ML models, between 0.651 and 0.861 (Table 7). RF and ANN had excellent performance in all performance indicators - all greater than or equal to 85%. However, ANN's recall rate is 74.9%, significantly lower than RF (85.8%), which led to a lower F1 score and AUC value than the Random Forest. Thus, RF yielded the best overall results.

### *5.2. Characteristics of value indicators of potential high-value patents*

RF had the best performance in recognition, and it identified 1812 potential high-value patents from a test set of 6292 patents with 95.3% accuracy. Table 8 shows the mean difference between HVPs and the entire sample and high-value and zero-cited patents. By examining the mean of patent indicators, HVPs belong to larger patent families (44.497), have longer patent instruction document (27.997), and cite more patents (30.267). There is little difference between the means of the claims and IPC classification indicators of HVPs, the entire sample (0.515 and 0.529), and the zero-cited patents (0.504 and 0.673). The more significant the difference between the means of HVPs' and the entire sample's patent indicators, the more prominent the legal, technological, competitive, and scientific value of high-value patents.

Nevertheless, the difference between the mean of HVPs' first citation speed indicator and the entire sample is negative. The larger the negative value, the better (−5.083). The mean differences imply that the RF model can distinguish potential high-value patents from the entire sample and the zero-cited patents.

Table 8 includes the mean multiple, which enabled us to compare the patent value indicators for the HVPs, the entire sample, and the zero-cited patents. The means of all HVPs' patent indicators are multiple times higher than that of the entire sample. A close examination suggests that non-patent references count (2.987), patent reference count (2.542), forward citations count (1.977), and extended patent family size (1.861) play a significant role in identifying PHVPs. As expected, the mean multiple of the first-citation speed indicator was less than 1. Compared to the mean difference, the mean multiple of claims (1.133) and IPC classifications (1.311) offer a stronger indication that the indicators helped identify high-value patents.

In general, the mean multiple of HVPs is higher than that of zero-cited patents. amongst all indicators, HVPs' mean of the non-patent reference count indicator is 6.197 times higher than the zero-cited patents' mean of the same indicator. The mean multiple also highlights the significant difference in patent family size (4.445) and patent reference count (3.945) between HVPs and zero-cited patents.

The comparisons of mean difference and mean multiple showed that patent family size, extended patent family size, length of patent instruction document, patent reference count, and non-patent reference count were good indicators to distinguish the HVPs

**Table 9**
The patent family number and their indicator values of some identifying high-value patents.

| Patent family number (only listing the first) | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| WO2015088486-A1 | 8 | 131 | 0 | 3 | 22 | 12 | 16 | 11 | 29 | 38 | 13 |
| WO2015048173-A2 | 13 | 132 | 3 | 7 | 21 | 11 | 10 | 2 | 148 | 13 | 10 |
| WO2015182000-A1 | 5 | 132 | 2 | 4 | 8 | 20 | 9 | 12 | 148 | 251 | 157 |
| US9100232-B1 | 5 | 128 | −3 | 2 | 104 | 14 | 3 | 7 | 38 | 50 | 9 |
| WO2015081882-A1 | 5 | 134 | 2 | 3 | 33 | 14 | 1 | 5 | 89 | 113 | 161 |
| WO2015033628-A1 | 5 | 131 | 2 | 2 | 29 | 17 | 2 | 8 | 65 | 45 | 17 |
| WO2015013245-A2 | 6 | 131 | 1 | 5 | 24 | 17 | 6 | 3 | 174 | 270 | 211 |
| WO2015056253-A1 | 5 | 132 | 3 | 8 | 15 | 13 | 3 | 5 | 65 | 91 | 75 |
| WO2015054567-A1 | 5 | 133 | 2 | 1 | 13 | 16 | 5 | 11 | 47 | 497 | 98 |
| WO2015176304-A1 | 5 | 135 | 1 | 3 | 12 | 17 | 4 | 4 | 49 | 65 | 40 |
| WO2015119198-A1 | 5 | 131 | 3 | 9 | 12 | 24 | 8 | 3 | 67 | 40 | 17 |
| WO2015155635-A1 | 8 | 132 | 1 | 3 | 11 | 12 | 1 | 4 | 79 | 196 | 75 |
| FR3021439-A1 | 5 | 141 | 2 | 7 | 10 | 20 | 6 | 7 | 26 | 32 | 18 |
| KR2015018358-A | 5 | 130 | 2 | 5 | 10 | 10 | 27 | 3 | 28 | 42 | 9 |
| WO2015199705-A1 | 5 | 136 | 3 | 1 | 9 | 15 | 9 | 4 | 36 | 39 | 12 |
| WO2015105924-A1 | 5 | 132 | 1 | 6 | 9 | 9 | 9 | 8 | 37 | 72 | 12 |
| WO2015183337-A1 | 5 | 135 | 1 | 4 | 7 | 13 | 4 | 6 | 50 | 173 | 78 |
| WO2015183377-A1 | 6 | 133 | 1 | 3 | 6 | 11 | 13 | 11 | 44 | 68 | 21 |
| WO2015017773-A1 | 9 | 131 | 5 | 6 | 48 | 16 | 7 | 4 | 32 | 24 | 21 |
| WO2015200021-A1 | 5 | 135 | 2 | 2 | 30 | 9 | 6 | 9 | 146 | 69 | 7 |

**Table 10**
Patent value indicators for each value dimension.

| Value dimension | Indicator name | Mean difference | | Mean multiple | |
|---|---|---|---|---|---|
| | | Entire sample | Zero-cited patents | Entire sample | Zero-cited patents |
| Legal | Patent family size | 27.894 | 34.487 | 1.595 | 4.445 |
| Technological | First-citation speed | −5.085 | N/A | 0.660 | N/A |
| Competitiveness | Forward citations | 1.707 | N/A | 1.977 | N/A |
| | Extended patent family size | 2.169 | 2.845 | 1.861 | 2.544 |
| Scientific | Length of patent instruction document | 11.027 | 13.510 | 1.650 | 1.933 |
| | Patent reference count | 18.361 | 22.594 | 2.542 | 3.945 |
| | Non-patents reference count | 4.766 | 6.008 | 2.987 | 6.197 |

from the rest. First-citation speed and forward citation also showed their prominent informative role in identifying HVPs. Table 9 contains examples of high-value patents that were identified by the study.

## 6. Conclusion and implications

### 6.1. Conclusion and discussion

The study modelled five ML based on a multidimensional patent value portfolio for automatic estimation. The patent value portfolio focused on patents' legal, technological, competitiveness, and scientific values. The random forest machine learning model had the best overall performance, and it identified a total of 1812 high-value IC patents. The results were robust (Table 7), suggesting that both the established patent value portfolio and the ML with a value portfolio pre-screening strategy developed in the study can effectively evaluate and identify PHVPs. The means of the patent indicators of the HVPs, the entire sample, and the zero-cited patents are sufficiently different, suggesting that the model tested in the study can distinguish the high-value patents from the rest (Table 8).

All value indicators in the portfolio were informative for high-value patent estimation, particularly patent family size, patent references, forward citations, and first-citation speed. Patent claims and IPC classifications are less critical in estimating HVPs in this study. For example, the difference between the mean of the claims indicator of HVPs and zero-cited patents is only slight (4.390 vs 3.886), and the mean multiple against the entire sample was only 1.133, and against patents receiving zero citations was 1.113. A similar result was also observed in IPC classifications, where the mean value of high-value patents is 2.232 compared to the mean value of the patents receiving zero citations, 1.559. However, it should be noted that using IPC classifications to measure patent value could be tricky because Morehrle and Frischkorn (2021) discovered that the narrow technological breadth could also lead to a high value not.

Table 10 identifies the patent value indicators in each value dimension playing a substantial role in evaluating and identifying potential high-value patents based on the mean difference and mean multiple.

The first citation speed indicator was used to estimate a patent's technological value. Based on immediacy, the patent citation speed mirrors the importance of technology and indicates how fast new knowledge is incorporated into the subsequent technological and scientific improvement (Bensen & Magee, 2015). Immediacy is a key characteristic that distinguishes the rapidly developing scientific fields from those that are not. Hence, more immediate use of patents leads to faster new knowledge incorporation and a

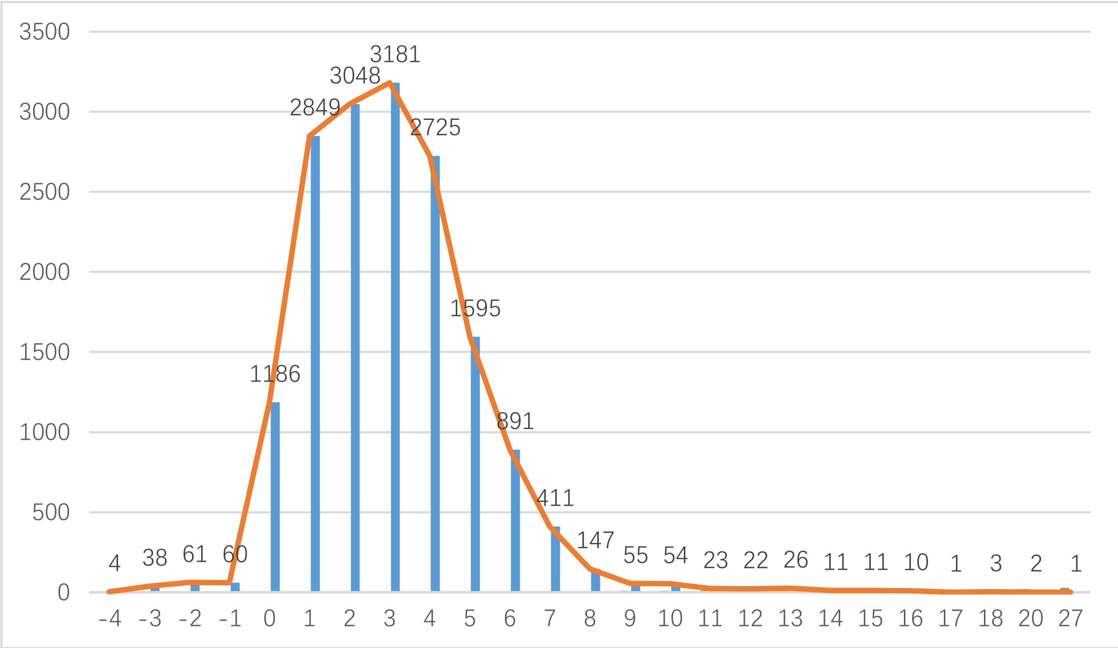


**Fig. 4.** Data distribution map of the first cited speed of patents.

**Table 11**
Performance indicators of five machine learning model classifiers based on strategy 1, strategy 2, strategy 3 and strategy 4.

| Classifier | Strategy 1 Accuracy | Precision | Strategy 2 Accuracy | Precision | Strategy 3 Accuracy | Precision | Strategy 4 Accuracy | Precision |
|---|---|---|---|---|---|---|---|---|
| NB Classifier | 0.848 | 0.754 | 0.866 | 0.545 | 0.885 | 0.24 | 0.912 | 0.343 |
| LR Classifier | 0.861 | 0.829 | 0.911 | 0.796 | 0.946 | 0.53 | 0.98 | 0.99 |
| RF Classifier | 0.953 | 0.978 | 0.992 | 1.000 | 0.944 | 0.477 | 0.906 | 0.223 |
| SVM Classifier | 0.942 | 0.924 | 0.970 | 0.902 | 0.945 | 0.517 | 0.988 | 0.991 |
| ANN Classifier | 0.942 | 0.942 | 0.991 | 0.953 | 0.945 | 0.5 | 0.958 | 0.673 |

higher rate of technological improvement (Bensen & Magee, 2015). The first-citation speed indicator potentially has negative values due to its advanced citation characteristic. The advanced citation occurs when a member of the same patent family has been cited before the publication of a given patent. Advanced citations reflect a higher technological value because the technology is derived from a patent family with better technical recognition and market performance. With this respect, HVPs' mean of the first-citation speed indicator of (Table 8) would be lower than the entire sample, and the mean multiple would be less than 1. Truncation refers to the truncation of patent citation time. Usually, different citation windows correspond to different citation numbers, potentially leading to significant deviations in experimental results. Therefore, most research uses short-term (5 years, 10 years) citations to solve this problem. The first citation speed is less affected by the length of the citation window, and most are first cited in 1 to 4 years (Fig. 4). Thus, the first-citation speed indicator effectively reduces the impact of the right truncation phenomenon on the research results.

We proposed a pre-screening strategy to pre-evaluate the top 5% HVPs with the multidimensional value indicators portfolio for constructing the positive target vector of training and testing. We conducted experiments using four strategies to test how robust and valuable the pre-screening approach is in constructing a positive target vector to identify PHVPs. Table 11 listed the ML models' performance in accuracy and precision using four pre-screening strategies.

- Strategy 1 is the most comprehensive and based on the entire multidimensional patent indicator portfolio,
- Strategy 2 uses four informative indicators, including patent family size, first cited speed, forward citations, non-patent reference count,
- Strategy 3 uses single-dimensional patent indicators such as patent family size, and
- Strategy 4 also uses single-dimensional patent indicators such as forward citations.

All four strategies were robust and effective. For example, when using the comprehensive pre-screening strategies 1 and 2, the models' performance in the accuracy range from 0.848 to 0.992; and from 0.885 to 0.988 when using single value dimension strategies such as 3 and 4. The results of strategies 1 and 2 give the comprehensive value of a patent which reflects different value dimensions, while the results of strategies 3 and 4 reflect the values of a single dimension.

### *6.2. Implications*

The study has implications for practice. First, the results suggest that our systematic approach for exploring and identifying the best-performing ML model for automatic HVPs identification is robust. Second, the proposed methodology process flow (Figs. 1, 2, & 3) presented in the study is a holistic approach and can be applied in the research on patent value estimation in other technological fields. Hence, firms and government agencies can use our approaches to establish standard procedures for the early identification of HVPs.

The study also has implications for research. First, the study explored different ML models using a pre-screening approach with a multidimensional value indicator portfolio to pre-evaluate HVPs. The strategy helps to lay down the foundation for the construction positive target vector later. Second, the study has demonstrated that the first-citation speed indicator is informative and valuable in value estimation. The indicator is an application of the concept of immediacy (Price, 1965) and added the time dimension to forward citations. The importance of the indicator should be recognized as an effective measure of patents' technological value more widely in the literature.

### *6.3. Further research*

Although RF yielded the best results with the given patent value portfolio, it should be noted that an ML model's performance can vary if different value indicators are used to construct the model or if the model is applied in a different field. It is because the degrees of informative value indicators for value estimation are not always the same in different areas (Haroff et al., 2003; Van Raan, 2017). Future research can be conducted to explore if RF developed in the study yields the same results in other fields. Moreover, future research can explore whether adding economic value indicators to the portfolio will change the outcomes of machine learning models. The study proposed the first-citation speed indicator and referred it to the speed of a patent that received its first citation. The indicator proved helpful in estimating patent value and distinguishing high-value patents in the study.

Nevertheless, the definition, scope, and calculation of first-citation speed can be developed further to ensure its robustness. Finally, this work collected the patent information from the Derwent database; therefore, the selections of the value indicators were constrained by the offerings of the database. Future work can utilize Derwent and Incopat databases to develop a more comprehensive indicator system for high-value patent estimation.

## Declaration of Competing Interest

No potential conflict of interest was reported by the author(s).

## Acknowledgement

This study was supported by the National Social Science Fund of China, "The Identification Method and Its Application in Identifying the "Hidden Treasures" from Massive Scientifical and Technical Literature" (Grant No. 20CTQ031).

## References

Albert, M. B., Avery, D., Narin, F., & McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy, 20*(3), 251–259.

Arts, S., Appio, F., & van Looy, B. (2012). Validating patent indicators that assess technological radicalness: The case of biotechnology. In *Proceedings of 17th International Conference on Science and Technology Indicators: 1* (pp. 82–97). Science-Metrix and OST.

Barney, J. A. (2010). *Statistic measurement of patent quality*. Ocean Tomo, LLC.

Benson, C. L., & Magee, C. L. (2015). Quantitative determination of technological improvement from patent data. *PloS One, 10*(4), Article e0121635.

Bessen, J. (2008). The value of U.S. patents by owner and patent characteristics[J]. *Research Policy, 37*(5), 932–945.

Breitzman, A., & Thomas, P. (2017). Inventor team size as a predictor of the future citation impact of patents[J]. *Scientometrics, 103*(2), 1–17.

Callaert, J., Looy, B. V., Verbeek, A., Debackere, K., & Thijs, B. (2006). Traces of Prior Art: An analysis of non-patent references found in patent documents. *Scientometrics, 69*(1), 3–20.

Caviggioli, F., De Marco, A., Montobbio, F., & Ughetto, E. (2020). The licensing and selling of inventions by US universities. *Technological Forecasting and Social Change, 159*, Article 120189.

Chen, D. Z., Lin, W. Y. C., & Huang, M. H. (2007). Using essential patent index and essential technological strength to evaluate industrial technological innovation competitiveness. *Scientometrics, 71*(1), 101–116.

Chi, Y. C., & Wang, H. C. (2022). Establish a patent risk prediction model for emerging technologies using deep learning and data augmentation. *Advanced Engineering Informatics, 52*, Article 101509.

Choi, J., Jeong, B., Yoon, J., Coh, B. Y., & Lee, J. M. (2019). A novel approach to evaluating the business potential of intellectual properties: A machine learning-based predictive analysis of patent lifetime. *Computers & Industrial Engineering, 145*, Article 106544.

Cricelli, L., Grimaldi, M., Rogo, F., & Strazzullo, S. (2021). Patent ranking indicators: A framework for the evaluation of a patent portfolio. *International Journal of Intellectual Property Management, 11*(2), 185–218.

Doonan, W., Higham, K. W., Governale, M., & Zülicke, U. (2019). Community structure in co-inventor networks affects time to first citation for patents. *Applied Network Science, 4*(1), 1–15.

Fasi, M. A. (2022). An Overview on patenting trends and technology commercialization practices in the university Technology Transfer Offices in USA and China. *World Patent Information, 68*, Article 102097.

Fisch, C., Sandner, P., & Regner, L. (2017). The value of Chinese patents: An empirical investigation of citation lags. *China Economic Review, 45*, 22–34.

Fischer, T., & Leidinger, J. (2014). Testing patent value indicators on directly observed patent value—An empirical analysis of Ocean Tomo patent auctions. *Research Policy, 43*(3), 519–529.
Frietsch, R., Schmoch, U., Van Looy, B., Walsh, J. P., Devroede, R., Du Plessis, M., & Schubert, T. (2010). The value and indicator function of patents (No. 15-2010). Studien zum deutschen Innovationssystem, Expertenkommission Forschung und Innovation (EFI) - Commission of Experts for Research and Innovation.
Frietsch, R., & Schmoch, U. (2010). Transnational patents and international markets. *Scientometrics, 82*(1), 185–200.
Gay, C., Le Bas, C., Patel, P., & Touach, K. (2005). The determinants of patent citations: An empirical analysis of French and British patents in the US. *Economics of Innovation and New Technology, 14*(5), 339–350.
Grimaldi, M., & Cricelli, L. (2020). Indexes of patent value: a systematic literature review and classification. *Knowledge Management Research & Practice, 18*(2), 214–233.
Grimaldi, M., Cricelli, L., Giovanni, M., & Rogo, F. (2015). The patent portfolio value analysis: A new framework to leverage patent information for strategic technology planning. *Technological Forecasting & Social Change, 94*, 286–302.
Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy, 32*(8), 1343–1363.
Hido, S., Suzuki, S., Nishiyama, R., Imamichi, T., Takahashi, R., Nasukawa, T., & Watanabe, T. (2012). Modeling patent quality: A system for large-scale patentability analysis using text mining. *Information and Media Technologies, 7*(3), 1180–1191.
Hou, J. L., & Lin, H. Y. (2006). A multiple regression model for patent appraisal. *Industrial Management & Data Systems, 106*(9), 1304–1332.
Hsieh, C. H. (2013). Patent value assessment and commercialization strategy. *Technological Forecasting & Social Change, 80*(2), 307–319.
Huang, C., Notten, A., & Rasters, N. (2011). and search strategies[J]. *Journal of Technology Transfer, 36*(2), 145–172.
Huang, Y., Chen, L., & Zhang, L. (2020). Patent citation inflation: The phenomenon, its measurement, and relative indicators to temper its effects. *Journal of Informetrics, 14*(2), Article 101015.
Huang, Z., Li, J., & Yue, H. (2022). Study on comprehensive evaluation based on AHP-MADM model for patent value of balanced vehicle. *Axioms, 11*(9), 481.
Kabore, F., & Park, W. G. (2019). Can patent family size and composition signal patent value? *Applied Economics, 51*(60), 6476–6496.
Kiehne, D. O., & Krill, B. (2017). The influence of the amount of inventors on patent quality. *Management, 47*, 300–308.
Kim, Y., Park, S., Lee, J., Jang, D., & Kang, J. H. (2021). Integrated survival model for predicting patent litigation hazard. *Sustainability, 13*(4), 1–15.
Kwon, U., & Geum, Y. (2020). Identification of promising inventions considering the quality of knowledge accumulation: A machine learning approach. *Scientometrics, 125*(3), 1877–1897.
Lai, K. K., Bhatt, P. C., Kumar, V., Chen, H. C., Chang, Y. H., & Su, F. P. (2021). Identifying the impact of patent family on the patent trajectory: A case of thin film solar cells technological trajectories. *Journal of Informetrics, 15*(2), Article 101143.
Lanjouw, J. O., & Schankerman, M. (2001). Characteristics of patent litigation: A window on competition. *The Rand Journal of Economics, 32*(1), 129–151.
Lee, J., & Sohn, S. Y. (2017). What makes the first forward citation of a patent occur earlier? *Scientometrics, 113*(1), 279–298.
Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change, 127*, 291–303.
Lin, H., Wang, H., Du, D., Wu, H., Chang, B., & Chen, E. (2018). Patent quality valuation with deep learning models. In *International conference on database systems for advanced applications* (pp. 474–490). New York: Springer.
Liu, W., Tan, R., Li, Z., Cao, G., & Yu, F. (2021). A patent-based method for monitoring the development of technological innovations based on knowledge diffusion. *Journal of Knowledge Management, 25*(2), 380–401.
Liu, W. D., Qiao, W. B., & Liu, X. (2021). Discovering the realistic paths towards the realization of patent valuation from technical perspectives: defense, implementation or transfer. *Neural Computing and Applications, 33*(2), 577–590.
Mehta, A., Rysman, M., & Simcoe, T. (2010). Identifying the age profile of patent citations: New estimates of knowledge diffusion. *Journal of Applied Econometrics, 25*(7), 1179–1204.
Moehrle, M. G., & Frischkorn, J. (2021). Bridge strongly or focus–An analysis of bridging patents in four application fields of carbon fiber reinforcements. *Journal of Informetrics, 15*(2), Article 101138.
Munari, F., & Oriani, R. (2011). *The economic valuation of patents: Methods and applications* (pp. XI–XVII). Edward Elgar Publishing.
Nagaoka, S. (2007). Assessing the R&D management of a firm in terms of speed and science linkage: Evidence from the US Patents. *Journal of Economics & Management Strategy, 16*(1), 129–156.
Noh, H., & Lee, S. (2022). Forecasting forward patent citations: Comparison of citation-lag distribution, tobit regression, and deep learning approaches. *IEEE Transactions on Engineering Management, 69*(4), 1185–1196.
Park, Y., & Park, G. (2004). A new method for technology valuation in monetary value: Procedure and application. *Technovation, 24*(5), 387–394.
Park, Y., & Yoon, J. (2017). Application technology opportunity discovery from technology portfolios: Use of patent classification and collaborative filtering. *Technological Forecasting & Social Change, 118*, 170–183.
Ploskas, N., Zhang, T., Sahinidis, N. V., Castillo, F., & Sankaranarayanan, K. (2019). Evaluating and ranking patents with multiple criteria: How many criteria are required to find the most promising patents? *Computers & Chemical Engineering, 123*, 317–330.
Reitzig, M. (2004). Improving patent valuations for management purposes—Validating new indicators by analyzing application rationales. *Research Policy, 33*(6–7), 939–957.
Ruan, X., Zhu, Y., Li, J., & Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP Neural Network. *Journal of Informetrics, 14*(3), Article 101039.
Ruan, X., Lyu, D., Gong, K., Cheng, Y., & Li, J. (2021). Rethinking the disruption index as a measure of scientific and technological advances. *Technological Forecasting and Social Change, 172*, Article 121071.
Seo, W. (2022). A patent-based approach to identifying potential technology opportunities realizable from a firm's internal capabilities. *Computers & Industrial Engineering, 171*, Article 108395.
Squicciarini, M., Dernis, H., & Criscuolo, C. (2013). *Measuring patent quality: Indicators of technological and economic value,* OECD science, technology and industry working papers. OECD Publishing *2013/03*.
Trappey, A., Trappey, C. V., Wu, C. Y., & Lin, C. W. (2012). A patent quality analysis for innovative technology and product development. *Advanced Engineering Informatics, 26*(1), 26–34.
Trappey, A. J., Trappey, C. V., Govindarajan, U. H., & Sun, J. J. (2019). Patent value analysis using deep learning models—The case of IoT technology mining for the manufacturing industry. *IEEE Transactions on Engineering Management, 68*(5), 1334–1346.
van Raan, A. F. (2017). Patent citations analysis and its value in research evaluation: A review and a new approach to map technology-relevant research. *Journal of Data and Information Science, 2*(1), 13–50.
van Zeebroeck, N., & Van Pottelsberghe de la Potterie, B. (2011). Filing strategies and patent value. *Economics of Innovation and New Technology, 20*(6), 539–561.
Wu, Y., Ji, Y., Gu, F., & Guo, J. (2021). A collaborative evaluation method of the quality of patent scientific and technological resources. *World Patent Information, 67*, Article 102074.
Wu, M. F., Chang, K. W., Zhou, W., Hao, J., Yuan, C. C., & Chang, K. C. (2015). Patent deployment strategies and patent value in LED industry. *PloS one, 10*(6), Article e0129911.
Xu, H., Winnink, J., Yue, Z., Zhang, H., & Pang, H. (2021). Multidimensional scientometric indicators for the detection of emerging research topics. *Technological Forecasting and Social Change, 163*, Article 120490.
Yang, S. Y., & Soo, V. W. (2012). Extract conceptual graphs from plain texts in patent claims. *Engineering Applications of Artificial Intelligence, 25*(4), 874–887.
Zhang, S. F., Yuan, C. C., Chang, K. C., & Ken, Y. (2012). Exploring the nonlinear effects of patent H index, patent citations, and essential technological strength on corporate performance by using artificial neural network. *Journal of Informetrics, 6*(4), 485–495.