Information Processing and Management 60 (2023) 103323

Contents lists available at ScienceDirect

Information Processing and Management

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

Identifying potentially excellent publications using a citation-based machine learning approach

Zewen Hu [a], Jingjing Cui [a], Angela Lin [b,*]

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

ARTICLE INFO

Keywords:
Machine learning
Artificial intelligence
Excellent papers
Highly cited papers
Sleeping beauty
Citation-based measures
Citation peak
Neural network
LightGBM
TabNet

ABSTRACT

Excellent research papers are vital to science and technology advances. Thus, the early identification of potentially excellent research papers and recognizing their value in science and technology is high on the research agenda. This study used a set of 5 static and 8 time-dependent citation features to explore six machine learning methods and identify the method with the best performance to identify potentially excellent papers. The study modelled Random Forest, LightGBM, Naive Bayes, Support Vector Machine, Neural Network, and TabNet to identify PEPs in the artificial intelligence field. The study defined highly cited papers using the threshold of the top 1% and top 5% and collected the data from the Web of Science®. Bibliometric and citation data from 485,041 research articles, proceeding papers, and reviews published in AI between 1990 and 2010 were collected initially. The data was screened and processed, and the final dataset consists of 96,169 papers for the training and test sets. The findings suggest that the time-dependent citation features are more important than the static features, and citation peak features are more significant than the citation features in identifying potentially excellent papers. The findings demonstrate the effect of threshold on machine learning outcomes (e.g., the top 1% and 5%); therefore, the study argues that the decision about threshold selection should be carefully made. LightGBM and Random Forest both performed with the given conditions and achieved the same score in accuracy and recall. Nevertheless, when comparing their performance in other indicators, such as F_1 and cross-entropy loss, LightGBM performed better. The study concluded that LightGBM was the best-performing model for identifying potentially excellent papers. The papers identified the contributions and recommended future research.

1. Introduction

Excellent research publications are vital in advancing knowledge and scientific and technological development. Therefore, the early identification of excellent publications is of interest to the research community, government agents, and firms. However, with the explosive growth in the number of research articles published each year, finding the existing excellent papers among many is already like looking for a needle in a haystack, not along with identifying potentially excellent publications (PEPs) at an early stage. This study wishes to contribute to the research in the early identification of excellent research publications.

What counts as an excellent article? There is not yet a consensus on what constitutes an excellent article in the existing literature.

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

https://doi.org/10.1016/j.ipm.2023.103323
Received 2 September 2022; Received in revised form 12 November 2022; Accepted 8 February 2023
Available online 16 February 2023

Much of the current literature considers that excellent articles are equivalent to those with higher citation counts; in other words, highly cited papers (HCPs) have been questioned despite using citation count to measure how good a paper is (Fu & Aliferis, 2010; Xie et al., 2019). The main concern is the quality and reliability of the measure because of the variation in citation behaviors (Tahamtan & Bornmann, 2018). Nevertheless, a comprehensive study by Bornmann et al. (2010) demonstrated that the papers contributing to scientific progress in a field are firmly based on previously important work. Bornmann et al. provided evidence to suggest that highly cited papers have higher scientific values and contribute more to scientific progress than those with fewer citations. Thus, this study defines excellent papers as those with high citation counts or highly cited papers (Bornmann, 2014). With this respect, the identification of PEPs involves citation prediction of papers.

Various factors contributing to citation count have been identified for citation prediction. Tahamtan et al. (2016) identified three categories covering 28 contributing factors to citation count: paper-, journal-, and author-related factors. Xie et al. (2019) identified 66 factors influencing citation counts and categorized them into four types: article-related, author-related, reference-related, and citation-related. Ruan et al. (2020) used 30 features relating to paper, journal, author, reference, and early citation to predict citation count. Xie et al. (2019) noticed that many studies relied on single-category citation features to model machine learning. They argued that even though some studies might have considered and used multi-category features to estimate citation count, they primarily emphasized the static features that do not change once an article is published. Time-dependent citation contributing factors are often overlooked in citation prediction studies. Based on this observation, this study answered the following questions as a part of the study:

- Do the static or time-dependent citation features have a more significant influence on citation prediction?
- Does a citation metric with static and time-dependent features have a more significant influence on the performance of citation prediction?

Given the research progress in artificial intelligence (AI), machine learning methods have become prominent methods for predicting and identifying highly cited papers (Akella et al., 2021; Liang et al., 2021; Lu et al., 2021; Ruan et al., 2020; Xu et., 2019). This, argued by Weihs and Etzioni (2017), is because the AI methods can produce more accurate long-term forecasts (e.g., ten years) compared with the traditional statistical modeling and manual approach. The studies deploying machine learning methods for citation prediction are primarily confirmatory. These studies are interested in exploring and confirming a specific learning method and provide little information about the performance of other methods in the same task. To address this issue, this study examined five baseline machine learning models (e.g., Random Forest, LightGBM, Naive Bayes, Support Vector Machine, Neural Network) and one recently emerged method (TabNet) to answer the following questions:

- Which machine learning model performs better with citation prediction with a given static and time-dependent citation features?
- Does the novel TabNet method designed for tabular data with interpretable canonical DNN learning architecture have a better performance in identifying PEPs?
- Can the threshold (e.g., top 1% and top 5%) for defining highly cited papers affect the performance of machine learning models?

This is an exploratory study, and its objectives are

- to examine the effect of the proposed citation metric on citation prediction
- to explore and identify the machine learning model with the best performance on citation prediction
- to investigate the impact of the definition of highly cited papers on the performance of the models

To achieve these objectives, we first established a set of features that comprises static features (e.g., article length, the number of keywords, the number of authors, abstract length, and the number of references) and time-dependent citation features (e.g., the first-citation speed, citations in the first year, citations in the first two years, fluctuation of annual citations, the number of citation peaks, the interval between citation peaks, the time between the first citation peak and publication year, and the time between the highest citation peak and publication year). We then modelled Random Forest, LightGBM, Naive Bayes, Support Vector Machine, Neural Network, and TabNet to identify PEPs in the AI field. 485,041 research articles, proceeding papers, and reviews published between 1990 and 2010 from the Web of Science® (WoS) were collected initially. The data was processed, and the final dataset consists of 96,169 papers with at least one citation peak. We labelled the papers ranked at the top 1% and 5% in the AI category in WoS published between 1990 and 2007 as the positive target vector of training and test sets. We used them as historical data to identify the potential excellent papers with higher citation value. The results yielded by the six machine learning models were then compared. LightGBM (Light Gradient Boosting Machine) was identified as having the best overall performance.

The contributions the study made to the existing literature are several. First, the study tested the classification performance of three kinds of machine learning models, including tree-based models (Random Forests, SVM, and LightGBM), Neural Network models (Neural Network, TabNet), and Probability-based models (Naive Bayes). The performances of these six models were compared, and LightGBM was identified as the best-performing model for identifying PEPs. This result differs from the studies that focused on exploring the strengths of neural network models in citation prediction. Thus, the study provides an alternative model for citation prediction. Second, the findings reveal the importance of time-dependent citation features, especially citation peak features, in identifying PEPs. These features were not in the previous research, and they are under-researched. Currently, there is not much research on citation peaks and their role in highly-cited papers and knowledge distribution. The finding of the importance of citation peak in identifying PEPs not only contributes but also raises awareness of the gap in the literature. Third, this study is one of the early

studies to deploy TabNet to identify PEPs. The findings demonstrate that the method is potentially useful. Fourth, the robustness test highlights the effect of a threshold for HCPs on the performance of machine learning. The results showed a trade-off between accuracy and recall using different thresholds (e.g., 1% and 5%). For example, the accuracy rate decreases but the recall rate increases when the threshold increases from 1% to 5% in the case of Random Forest, LightGBM, SVM, Neural Network and TabNet. In contrast, the accuracy rate decreases, and the recall rate also decreases when the threshold increases from 1% to 5% in the case of Naive Bayes.

We organized the article into seven distinct sections. The next section presents a literature review of the related work on citation prediction and machine learning methods for citation prediction. Section 3 explains the study's research design, including data collection methods and the six machine learning models used for identifying PEPs. Section 4 details the implementations of machine learning models, and Section 5 presents the results. Section 6 presents the results of the robustness test of the machine learning models using a different threshold. The article with a discussion and recommendations for future studies.

2. Related work

2.1. Features associated with highly cited papers(HCPs)

2.1.1. Static features

Citation count is affected by an array of factors. The factors can be categorized into the article, author, journal, reference, content, and citation-related features. The factors can further be categorized into static, dynamic, and time-dependent features according to their temporal characteristics. Static features are intrinsic to an article; once an article is published, they cannot be changed, for example, title, abstract, author(s), affiliations, keywords, and references. Dynamic features vary according to the situation post-publication, such as citation network. Time-dependent features change over time post-publication. The static features commonly used for citation predictions include the length of abstracts, the number of authors, the number of keywords, the length of an article, and the number of references. These features have been proven positively and significantly associated with citation count in the current studies.

Citation count increases as the number of authors increases (Falagas et al., 2013; Robson & Mousquès, 2016; Wendzel et al., 2020). Aksnes (2003) explained that papers with many authors receive higher citations because they benefit from potential self-citations and enhanced dissemination through their networks in the community. Robson and Mousquès (2016) assessed 6122 environmental modeling papers to identify bibliometric and categorical variables related to citation counts and discovered that citation count increases as the number of authors increases. An analysis of citations in information security research (Wendzel et al., 2020) suggests that a paper with 2–7 authors performs better in citation count. This supports an earlier study (Iqbal et al., 2019) which indicates that "most-cited articles typically have a moderate number of authors" in the domain of computing networking (e.g., 2–7). Wendzel et al. (2020) found that this effect vanishes when the number of authors exceeds a threshold of 8. They explained that this could be because a paper may be perceived as lacking credibility if there are too many authors.

The number of keywords positively correlates with citation counts (Fiala et al., 2021; So et al., 2015; Uddin & Khan, 2016; Wendzel et al., 2020). This is because a keyword can convey important information about the research. The higher the number of keywords, the broader range of subject matters being covered, which increases the chance that the paper reaches its target audience (Uddin & Khan, 2016).

HCPs tend to have more extended abstracts than non-HCPs (Hafeez et al., 2019; Robson & Mousques, 2016; Wendzel et al., 2020). Robson and Mousquès (2016) found that citation count increases smoothly if the abstract's word count increases; this effect reaches its maximum of around 300 words. Wendzel et al. (2020) explained that more extended abstracts could include more and longer keywords which enable free-text searches. Therefore, the articles are more likely to be discovered by search engines than those with shorter abstracts.

Article length, measured by the number of pages, is positively associated with the citation counts (Falagas et al., 2013; Hafeez et al., 2019; Lyu & Wolfram, 2018; Robson & Mousquès, 2016; So et al., 2015; Vanclay, 2013; Wendzel et al., 2020). For example, a study of the top five economics journals and their citation counts in Google Scholar discovered that a 1% increase in page length is associated with a 0.56% increase in the number of citations (Hasan & Breunig, 2021). Falagas et al. (2013) argued that article length could independently predict citation counts because "the greater article length could reflect increased greater scientific complexity and higher methodological quality of a study; in addition, lengthier articles are expected to contain more information, thus increasing the possibilities that part of it will be appropriate to be cited by other researchers" (p.4).

An article with a higher number of references is usually associated with a higher citation rate (Didegah & Thelwall, 2013; Falagas et al., 2013; Hafeez et al., 2019; So et al., 2015; Wendzel et al., 2020; Xie et al., 2019). This could be because references made the work more visible via citation-based search (Didegah & Thelwall, 2013). In addition, Wendzel et al. (2020) argued that the number of references indicates that the work is grounded in the existing literature and demonstrates its authoritativeness and thoroughness.

2.1.2. Time-dependent citation features of highly cited papers

Time-dependent citation features display HCPs' influence on subsequent research over time. Evidence has shown that excellent papers have longer citation life and faster dissemination speed than papers with fewer citations (Aversa, 1985; Glänzel & Garfield, 2004; Aksnes, 2003). Aversa (1985) studied the relationship between citation patterns and the literature aging rate of 400 HCPs and identified two distinct citation patterns: *delayed rise, slow decline,* and *early rise, rapid decline*. The aging rate of the former is slower than the latter. Ponomarev et al. (2014) studied the pattern of time-dependent citation behavior of HCPs. They found that papers, including top-ranking papers, experience an initial period of slow citation growth, which lasts 5 to 20 months. Following this initial slow growth

period, the citation rates accelerate, reach saturation plateaus, and then decline. However, 25% of the top-cited papers continue experiencing the first stage of citation growth, 50% reach saturation point, and 25% start falling after 5 years. Ponomarev et al. constructed forecasting models for predicting highly cited papers based on these citation patterns.

Time-dependent citation features have also been used to study and identify the characteristics of sleeping beauties, especially the delay-recognition feature. Sleeping beauties have *delayed rise and slow decline* characteristics usually go unnoticed and receive no or very few citations in the following years of publication (Van Raan, 2004). They only start receiving citations later (Lachance & Larivière, 2014). Once a 'sleeping beauty' is awakened by its first citing paper, termed a prince, it accumulates many citations. Sleeping beauties are rare, often found in natural science, and least in social science (Li & Ye, 2014). They tend to involve discoveries or new concepts which require time for people to accept them or for the knowledge to be fully integrated into the field (Lachance & Larivière, 2014; Ohba & Nakao, 2012). Although sleeping beauties receive recognition late, they may substantially influence later research and scientific breakthroughs. Highly cited papers could include sleeping beauties.

Based on the above, time-dependent citation features are important indicators for citation prediction. However, only a few studies have systematically incorporated citation features for citation prediction (Abramo et al., 2019; Ruan et al., 2020; Xie et al., 2019; Yu et al., 2014). Xie et al. (2019) identified seven time-dependent citation features, including first cited age, citation count in the first year, citation count in the first 2 years, citation count in the first 5 years, number of citations citing journals in the first year, number of citations citing journals in the first 2 years, and number of citations citing journals in the first 5 years. First cited age (Yu et al., 2014) or first-citation speed, the term used in this study, refers to the time lag between a paper receiving its first citation since publication. The first citation is significant because it is the turning point where a paper changes its status from being uncited to cited. First-citation speed indicates the speed of knowledge diffusion in the community and is argued to correspond positively to the impact of a paper (Yu et al., 2014). Most articles tend to receive their first citation within the first three years of publication, and papers that are cited beyond three years or more are considered to have delayed-recognition status (Lachance & Larivière, 2014; Li & Ye, 2016). Xie et al. (2019) discovered that all citation-related features strongly correlated with the total citations. Yu et al. (2014) found that the first-cited age and citation count after the first 2 years of publications are good indicators for paper citation impact. Ruan et al. (2020) borrowed most citation-related features to form a set of early citation features for citation prediction. The findings support the crucial role of early citations, especially 'citations in the first 2 years' and 'first-cited age' in citation prediction (Abramo et al., 2019; Yu et al., 2014).

Including time-related citation features in the citation prediction can reduce the risk of overlooking sleeping beauties. Hence, time-related citation features associated with sleeping beauties can be added to the existing list of citation features (Ruan et al., 2020; Xie et al., 2019). These time- and sleeping beauty-related features are the fluctuation of annual citation count, the number of citation peaks, the interval between citation peaks, the time between the first citation peak and publication year, and the time between the highest citation peak and publication year.

A paper's annual citation count indicates how influential and attractive a paper is in its research field. Annual citations fluctuate, and the fluctuation frequency reflects the increase or decrease in attention that an article receives. Fluctuation range of annual citations concerning the deviation of the yearly citations from the average citations within a specific period since publication. Low values indicate little change in the degree of attention received and vice versa (Li & Ye, 2014).

A citation peak is "the highest number of received citations, starting from the year of publication" (van Leeuwen, Moed & Reedijk, 1999, p. 494). The speed to reach the first citation peak indicates the speed of knowledge diffusion. A vast majority of papers' citation counts reach their first peak a few years after publication, e.g., 1–2 years (Li et al., 2019; van Leeuwen et al., 1999, 2003), but some may peak later (Wang et al., 2013; Li & Ye, 2014). Citation peaks are a prominent indicator of citation pattern analysis (Li et al., 2019). As put, "the peaking time of citations features the shape of citation curves, reflecting immediacy of publications" (Li & Ye, 2014, p.19). The highest citation peak refers to the peak with the most significant peak value among all peaks, and most articles reach the highest citation peak within 0–5 years of publication (Li et al., 2019). Li et al. (2019) found that most of the papers (62.7%) had only one citation peak, and only a few papers with two and more peaks (e.g., 25.9% with two peaks, and 7.4% with three peaks). They (2019) further argued that the interval between citation peaks of a paper is meaningful. For example, when the interval(s) between peaks is short, the article receives intensive attention in a specific timeframe. When the interval(s) between peaks is long, the paper may experience two lifecycles or is a sleeping beauty.

2.2. Methods for citation prediction

The regression approach is commonly used for citation prediction. Falagas et al. (2013) used a backward multiple linear regression model to identify independent predictors of the higher number of citations and a multiple regression model with logarithmic transformation of the dependent variable (citation count) to assess logarithmic. They studied the prediction power of 8 variables comprising the number of authors and affiliated institutions, title and abstract word count, article length, number of references, study design, access to the article (open access or requiring a subscription), and 2006 journal impact factor (JIF). They found only the article length and JIF independently predicted citation count. Yu et al. (2014) argued that the prediction accuracy of the existing prediction methods was not satisfactory and proposed using more objective features of scientific papers. Yu et al. identified 24 objective features using a multiple stepwise regression model to identify good variables from all the features for a model describing the relationship between the features and citation impact. They argued that the prediction of paper citation impact after 5 years of publication in Information Science & Library Science is improved with relative accuracy. Abramo et al. (2019) used a combination of early citation and journal impact factors to predict papers' long-term impact. They discovered that a three-year citation window with a linear regression model could sufficiently predict the long-term impact of scientific publications.

Machine learning methods have advantages in autonomous learning, probabilistic learning, feature matching, value classification,

and recognition prediction. Research using machine learning methods for citation prediction usually involves large datasets and produces more accurate long-term forecasts (e.g., ten years) (Weihs & Etzioni, 2017). Machine learning methods have been applied to predict scholarly impact (Akella et al., 2021; Xu et al., 2019), citation count, and h-indices (Huang et al., 2022; Mistele, Price & Hossenfelder, 2019; Weihs & Etzioni, 2017), research topics and topic trends (Liang et al., 2021; Lu et al., 2021), and to distinguish between sleeping- and non-sleeping beauties in the literature (Dey et al., 2017). The machine learning methods include SVM, Random Forest, K-nearest neighbor (kNN), Gaussin progress regression (GPR), and Classification and Regression Tree (CART). Fu and Aliferis (2010) used SVM to predict the citation count of biomedical papers published from 1991 to 1994, setting the citation window as ten years. The AUC of the model ranged between 0.86 and 0.92, while an AUC of 0.80 indicates a highly predictive classifier. This result was validated by comparing it with the logistic regression-based classifier. Robson and Mousquès (2016) applied the Random Forest model to predict citations of environmental modeling papers. They chose the model for its flexibility and not assuming a linear response or normally distributed response variable. Robson and Mousquès argued that the model allowed them to separate the cumulative effects of input variables to examine each variable's impact on citation prediction.

Neural network learning methods are widely used for citation prediction (Xu et al., 2019). A neural network simulates the structure and function of a neural network in the human brain for information processing. An advantage of neural network methods is that data distribution is not required, unlike the traditional regression approach. Studies using neural network learning methods have reported satisfactory outcomes of citation prediction. Evidence suggests that BP neural network model was significantly better than methods based on linear regressions (Lee & Choeh, 2014; Wong & Chan, 2015; Ruan et al., 2020; Wong et al., 2017). Ruan et al. (2020) used BP neural network to predict citations of academic papers. They selected 49,834 articles published in the Library, Information, and Documentation field from 2000 to 2013 as their dataset. Ruan et al. compared the performance of the BP neural network with six baseline models and demonstrated BP neural network's superiority in citation prediction. Liang et al. (2021) employed deep learning neural networks with nine bibliographical features to predict research trends, and keywords were one of the features. The results showed that LSTMA and NNAR outperformed rule-based naïve models. Xu et al. (2019) used a covolutional neural network (CNN) and bibliographic features to predict the scientific impact of papers. CNN was adopted to address the temporal evolution of the features that other machine learning methods do not address. The findings reveal that papers receiving attention from the community at an early stage have a long-term impact on subsequent research. Xu et al. (2019) provided evidence for the claim that their model yielded better prediction than some existing studies and demonstrated that CNN is superior to the Support Vector Machine (SVM) and Multiple Linear Regression model. Yuan et al. (2018) proposed and used a deep learning attention mechanism (DLAM) to consider papers' ability to attract attention as a feature to forecast citations. The experiment results suggest that the longer the training set duration, the better the longer-term prediction compared with other methods. For example, in t1, DLAM's MAPE and ACC performance were only slightly better than RNN, 2.13% and 1.65% higher, respectively. In t5, DLAMs performance in MAPE and ACC was significantly superior to RNN, 24.31% and 16.7%, respectively.

While neural network methods are proven helpful for citation prediction, the methods rely on much training data for an accurate forecast. For example, Du et al. (2022) found that the lowly cited papers accounted for the vast majority; therefore, using the trained LSTM network can capture accumulative patterns of the citation counts of those papers better and achieve more acceptable results on the entire dataset compared to its performance on highly cited articles. Abrishami et al. (2019) also pointed out that numerous technical challenges must be overcome despite the deep learning models having better prediction capability and producing more accurate predictions (Abrishami & Aliakbary, 2019). For example, while deep learning has been successful with image, text, and audio data, it is still not successful with tabular data (Arik & Pfister, 2021). TabNet, a canonical Deep Neural Networks (DNN) architecture for

Table 1
The descriptions of the measurements used for constructing feature vector space.

Number	Feature	Description
Static features		
X_1	Number of authors	Number of authors for the paper
X_2	Number of keywords	Number of the author-defined keywords in the paper
X_3	Abstract length	Word count of the abstract.
X_4	Paper length	Number of pages in the article
X_5	Number of references	Number of references cited in the paper
Time-dependent citation features		
X_6	First-citation speed	The time lag between a paper receiving its first citation since publication.
X_7	Citations in the first year	The number of citations in the first year after publication.
X_8	Citations in the first 2 years	The number of citations in the first two years after publication.
X_9	Fluctuation of annual citation	The fluctuation in annual citations within the citation cycle. Calculation: The value is presented as the ratio of the standard deviation of the citation frequency to the mean of the citation frequency in each paper's entire life citation cycle.
	Citation peak features	
X_{10}	Number of citation peaks	The number of citation peaks reflects a paper's use-value or citation value.
X_{11}	The interval between citation peaks	The period between two citation peaks.
X_{12}	The time between the first citation peak and the publication year	The interval between a paper's publication year and its first citation peak.
X_{13}	The time between the highest citation peak and publication year	The interval between a paper's publication year and its highest citation peak.

tabular data, was proposed for the problem (Section 3).

3. Methods

3.1. Features and the values

The static and time-dependent citation features used in this study are summarized in Table 1. The information on the features can be downloaded from the WoS directly, and the frequencies were counted using a Python program. The values of time-dependent citation features were calculated using the citation distribution data. The formulas used for calculating the values of citation peak features are detailed below (Zhao & Li, 2015; Yoon, 2017).

Given C_t is the citation frequency of a paper in the year t following publication. $C_t$is regarded as a citation peak if and only if C_t of an article is a local maximum, and its value is greater than c of the satisfying formula (1). c is the constant threshold, and Δt is a specific citation period within the entire citation window.

$$C_t > \max\{C_{t-\Delta t}, \ldots, C_{t-1}\} \text{and} C_t \geq \max\{C_{t+1}, \ldots, C_{t+\Delta t}\}, \text{and} C_t > c \tag{1}$$

This study took Δt=3, $c = 2$, so a given citation period of 7 years is (2Δt+1) if the citation frequency of a publication in the year t satisfies formula (2). The point (t, C_t) is considered a citation peak of the article. Formula (2) was for calculating the Number of citation peaks, the Interval between citation peaks, the Time between the first citation peak and publication year, and the Time between the highest citation peak and publication year. The intervals between multi-citation peaks of a paper are the average value of all the values of two adjacent peaks added together.

$$C_t > \max\{C_{t-3}, \ldots, C_{t-1}\} \text{and} C_t \geq \max\{C_{t+1}, \ldots, C_{t+3}\}, \text{and} C_t > 2 \tag{2}$$

3.2. The design of machine learning classifiers

The study explored Random Forest, LightGBM, Naive Bayes, Support Vector Machine, Neural Network, and TabNet classifiers for citation prediction and identification of PEPs in the AI.

(1) **Random Forest**

The classification processes based on Random Forest are:

Step 1: Randomly choose s ($s<X$) samples in the sample set and randomly select t ($t<Y$) bibliometric characteristics from all features. Suppose there are X samples and Y features in the sample set.
Step 2: Use t bibliometric characteristics of s samples to train the decision tree model.
Step 3: Repeat steps 1 and 2 m times to generate m decision trees to form a Random Forest.

For new samples, through m decision-making trees, the final classification result will be determined based on a majority voting mechanism.

(2) **Light Gradient Boosting Machine (LightGBM)**

LightGBM is a Gradient Boosting Decision Tree (GBDT) algorithm proposed by Microsoft in 2017 (Ke et al., 2017). The essence of GBDT is to upgrade a weak learner to a more robust learner. This is achieved by using the error rate of the previous iteration of the weak learner to update the weight of the training set. As follows, taking the negative binomial log-likelihood loss function as an example (formula (3)), the algorithm flow of GBDT binary classification is introduced (Friedman, 2001).

$$L(y, \mathrm{F}) = \log(1 + \exp(-2yF)), y \in \{-1, 1\} \tag{3}$$

Among them, y is the true value, and F is the prediction function, $F = \frac{1}{2}\log\left[\frac{P(y=1|x)}{P(y=-1|x)}\right]$. It can be proved that formula (4) is consistent with the logistic regression loss function, and the proof step is omitted:

$$L(y, F) = y\log(P(y=1|x)) + (1-y)\log(1 - P(y=1|x)), y \in \{0, 1\} \tag{4}$$

The algorithm flow of GBDT binary classification:

Step 1: $F_0(x) = \mathrm{argmin}_\rho \sum_{i=1}^{N} L(y_i, \rho)$, N represents the sample size, $y_i$represents the true value, ρ represents the initial base learner.
Step 2: For m = 1,2,3,…, M, to generate M weak learners:

(a) GBDT minimizes the loss function by iteratively creating a weak learner that points in the direction of the steepest descent (negative gradient direction).

(b) The leaf node area corresponding to the learned m^{th} regression tree is $\{R_{jm}\}_{1,}^{J}$ $j = 1,\ldots,$J. Each base learner is a J-terminal regression tree.

(c) The regression tree was used as the base learner to obtain the optimal leaf node value (formula (5)):

$$\gamma_{m,j} = argmin_{\gamma} \sum_{x \in R_{m,j}} \log(1 + \exp(-2y(F_{m-1}(x) + \gamma))) \tag{5}$$

(d) The strong learner obtained in this iteration (formula (6)):

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J} \gamma_{m,j} I\left(x \in R_{m,j}\right) \tag{6}$$

Step 3: The strong learner was obtained by integrating M weak learners (formula (7)):

$$F_M(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} \gamma_{m,j} I\left(x \in R_{m,j}\right) \tag{7}$$

Step 4: Finally, the probability estimation was performed according to the obtained prediction function, and the samples were classified according to the probability (formula (8)):

$$\begin{aligned} P(y = 1|x) = p = \frac{e^{2F(x)}}{1 + e^{2F(x)}} = \frac{1}{1 + e^{-2F(x)}} \\ P(y = -1|x) = 1 - p = \frac{1}{1 + e^{2F(x)}} \end{aligned} \tag{8}$$

The single-sided gradient sampling algorithm (GOSS) and the mutually exclusive feature bundling algorithm (EFB) in LightGBM were used to reduce the number of samples and bibliometric variables, respectively, thereby reducing the complexity of the model splitting process and ensuring accuracy.

(3) **Naive Bayes**

The Naive Bayes classification algorithm identifies EPs by calculating the probability that a publication belongs to the excellent or ordinary paper category.

Step 1: Construct a probability vector $(w_1, w_2, \ldots, w_k, \ldots, w_n)$ and use formula (9) to calculate the probability of each feature item of the publications belonging to each category. $N(W_k, d_i)$ is the feature value of W_k in d_i, and n is the total number of feature items.

$$w_k = P\left(W_k \middle| C_j\right) = \frac{1 + \sum_{i=1}^{|D|} N(W_k, d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N(W_s, d_i)} \tag{9}$$

Step 2: Using formula (10) to calculate the probability of a new publication d_i belonging to category C_j.

$$P\left(C_j|d_i;\widehat{\theta}\right) = \frac{P\left(C_j|\widehat{\theta}\right)\prod_{k=1}^{n} P\left(W_k|C_j;\widehat{\theta}\right)^{N(W_k,d_i)}}{\sum_{r=1}^{|C|} P(C_r|\widehat{\theta})\prod_{k=1}^{n} P(W_k|C_r;\widehat{\theta})^{N(W_k,d_i)}} \tag{10}$$

Above:

$$P\left(C_j|\widehat{\theta}\right) = \frac{\text{numberof}\, C_j - \text{typetrainingpapers}}{\text{Totalnumberoftrainingpapers}},$$

The calculation of $P(C_r|\widehat{\theta})$ is the same as of$P(C_j|\widehat{\theta})$ but replacing the values of C_j withC_r in the calculation $|C|$ is the total number of classes

Step 3: Comparing the probabilities of the new article d_i belonging to category C_j, and then classifying the article into the category with the highest probability.

The Naive Bayes algorithm assumes that under the given categorical variables, all feature items are independent of each other, but this assumption is unrealistic in practice. Thus, the classification results based on the algorithm may be less than ideal.

(4) **Support Vector Machine**

The Support Vector Machine (SVM) classifier identifies EPs using two approaches. The first approach is constructing an optimal classification hyperplane for the linear division of EPs and non-EPs. The optimal classification hyperplane enables feature vectors in training set to be linearly segmented without error. It maximizes the distance between the feature vectors closest to both sides of the hyperplane. The SVM is a binary classifier; therefore, when the samples are linearly inseparable, the Gaussian kernel function was used to map each sample point to an infinite-dimensional feature space, thereby making the linearly inseparable data linearly separable. Function K (x, z) is the kernel function, including the linear kernel, polynomial kernel, and Gaussian kernel, and Φ is a mapping from low-dimensional feature space to high-dimensional feature space (formula (11)).

$$\mathrm{K}(x,z) = \phi(x) * \phi(z) \tag{11}$$

(5) **Neural Network**

The Neural network model can automatically learn appropriate weight parameters from data. The model minimizes the loss function value by continuously updating the weights. The neuron of the Neural Network's input layer accepts the feature vector X of publications as the external input, and the hidden layer accepts the linearly weighted X, namely *a*. The hidden layer undergoes a nonlinear change and enters the next hidden layer. Taking the 2-layer neural network as an example, vector a represents the state of the hidden layer (Fig. 1).

$$\begin{aligned} a_1^{(2)} &= f\left(W_{11}^{(1)}x_1 + W_{12}^{(1)}x_2 + W_{13}^{(1)}x_3 + b_1^{(1)}\right) \\ a_2^{(2)} &= f\left(W_{21}^{(1)}x_1 + W_{22}^{(1)}x_2 + W_{23}^{(1)}x_3 + b_2^{(1)}\right) \\ a_3^{(2)} &= f\left(W_{31}^{(1)}x_1 + W_{32}^{(1)}x_2 + W_{33}^{(1)}x_3 + b_3^{(1)}\right) \\ h_{W,b}(x) &= a_1^{(3)} = f\left(W_{11}^{(2)}a_1^{(2)} + W_{12}^{(2)}a_2^{(2)} + W_{13}^{(2)}a_3^{(2)} + b_1^{(2)}\right) \end{aligned} \tag{12}$$

In formula (12), W is the weight, b is the constant term, and h is the output value of the previous layer (the input value of the next layer). Fig. 2. is the Neural Network model framework used in the study. The output layer combined with the sigmoid function to realize the two classifications: PEPs and non-EPs.

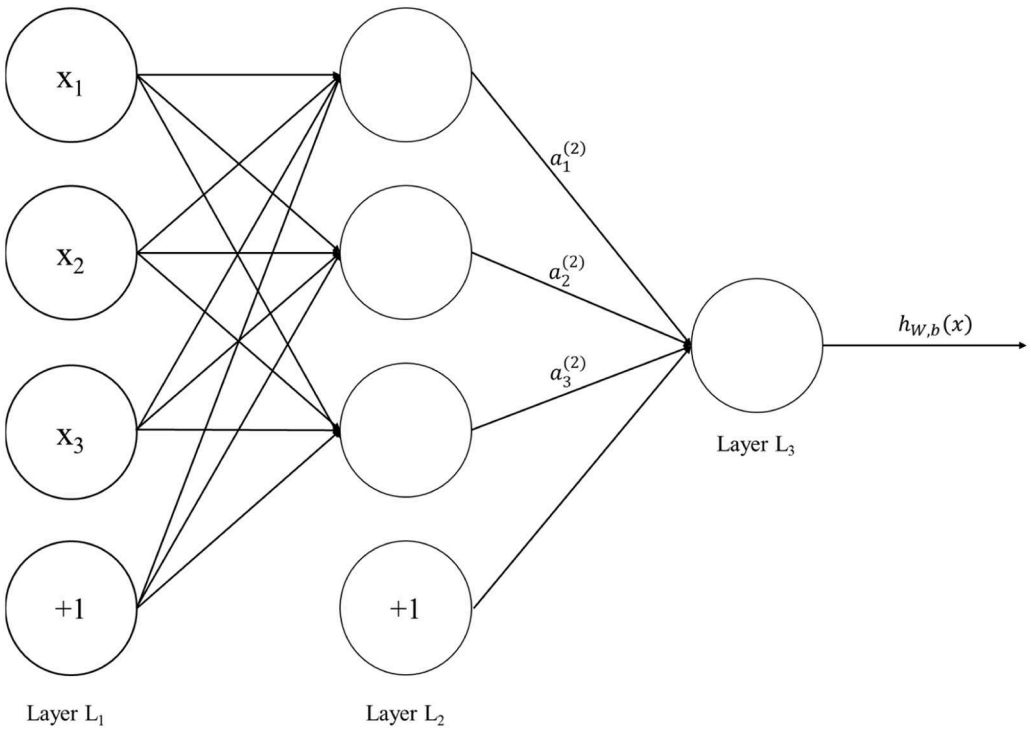

Fig. 1. A two-layer Neural Network model.

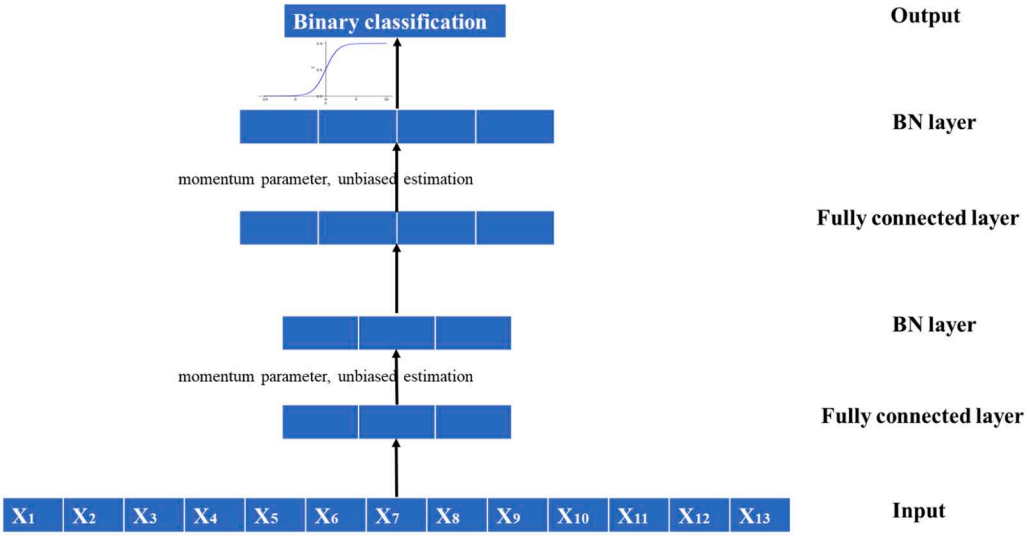

Fig. 2. The Neural Network model.

(6) **TabNet**

TabNet is a novel high-performance and interpretable canonical DNN learning architecture for tabular data (Arik & Pfister, 2021). The benefits of TabNet outperforming Decision Tree-based models are as follows: (1) TabNet inputs raw tabular data without any pre-processing and uses sparse instance-wise feature selection learned from data to choose a subset of semantically meaningful features; (2) constructs a sequentially attentive multi-step architecture to obtain decision boundaries in hyperplane form in each decision step based on the selected features; (3) improves the learning capacity via nonlinear processing of the selected features. Finally, a decoder-encoder architecture to realize self-supervised learning is designed, as shown in Fig. 3, composed of a feature transformer, an attentive transformer, and feature masking.

The decoder-encoder process of implementing TabNet for self-supervised learning of classifying PEPs and non-Eps is as follows:

Step 1 (Inputting and Passing Features): The input is B * D, B is the batch size, D is the dimension of the features, $f \in R^{B\times D}$ represents the D-dimensional features.When passing the same D dimensional features $f \in R^{B\times D}$ to each decision step. TabNet's encoding is based on sequential multi-step processing with N_{steps} decision steps.

Step 2 (Normalizing and processing features): Implementing the batch normalization of features through the BN layer and processing the filtered features through the feature transformer layer. Then split for the decision step output and information for the subsequent step.

Step 3 (Feature selection): employing a learnable mask M[i] $\in R^{B\times D}$ for soft selection of the salient features. Through the sparse selection of the most salient features, the learning capacity of a decision step is not wasted on irrelevant ones.

Step 4 (Split layer): The split layer cuts the vector output from the previous layer into two parts. One part calculates the final output of the model, and the other calculates the next mask layer.

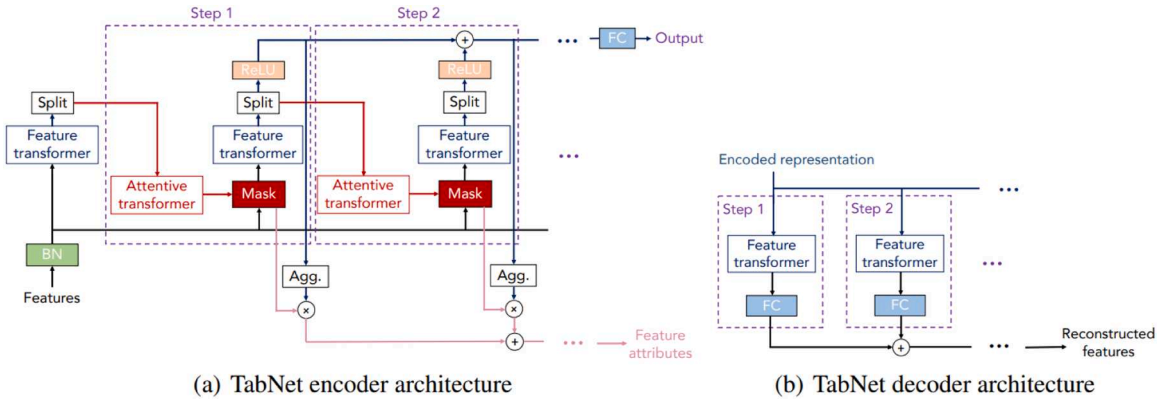

Fig. 3. The basic decoder-encoder architecture of TabNet for self-supervised learning.

Table 2
Confusion matrix.

	True results	Predicting results
	Positive (P)	Negative (N)
Positive (P)	TP	FN
Negative (N)	FP	TN

Step 5 (Transformer layer): The attention transformer layer is for feature selection, which calculates the current mask layer according to the previous step's results.

$$\mathrm{M[i]} = \mathrm{sparsemax}(\mathrm{P}[i-1] \cdot \mathrm{h_i}(\mathrm{a}[i-1])) \tag{13}$$

Where M[i] is a learnable mask for selecting the salient features, P[i] is the prior scale term, denoting how much a particular feature has been used previously, and $\mathrm{h_i}$ is a trainable function.

$$L_{sparse} = \sum_{i=1}^{N_{steps}} \sum_{b=1}^{B} \sum_{j=1}^{D} \frac{-M_{b,j}[i]\log\left(M_{b,j}[i]+\varepsilon\right)}{N_{steps} \cdot B} \tag{14}$$

ε is a small number for numerical stability, $\sum_{j=1}^{D} M[i]_{b,j} = 1$.

Step 6: Feature attribute aggregates the global importance of features.

$$M_{agg-b,j} = \sum_{i=1}^{N_{steps}} \eta_b[i] M_{b,j}[i] \Big/ \sum_{j=1}^{D} \sum_{i=1}^{N_{steps}} \eta_b[i] M_{b,j}[i] \tag{15}$$

$\eta_b[i] = \sum_{c=1}^{N_d} ReLU(d_{b,c}[i])$ is the aggregate decision contribution at $\mathrm{i^{th}}$ decision step for the $\mathrm{b^{th}}$ sample.

Step 7 (Output): The output of the categorized task is a vector.
Step 8 (TabNet decoder): TabNet decoder, composed of a feature transformer block at each step. The decoder architecture reconstructs tabular features from the TabNet encoded representations.

3.3. Evaluation of performances of the machine learning models

The study selected six widely used indicators in data mining and machine learning to evaluate the performances of our machine learning models, including accuracy, precision, recall, F1, P-R curve, and two-class cross-entropy loss. The study used high-impact articles as positive examples and low-impact publications as counterexamples to calculate accuracy, recall, and F1 (Table 2). Table 3 summarises the definition and calculation of the evaluation indicators.

The two-class cross-entropy loss function is derived from the maximum likelihood estimation. For a data set with N samples, the average loss on all samples is written as:

$$L(\omega) = -\frac{1}{N} \sum_{i=1}^{N} (y_i * \ln(\sigma_i) + (1-y_i) * \ln(1-\sigma_i)) \tag{16}$$

Table 3
The description of the indicators for evaluating the performance models.

Indicator	definition	Calculation formula
Accuracy	The proportion of correctly classified samples in the total sample	$\frac{TP+TN}{TP+TN+FP+TN}$
Recall	The proportion of the high-impact publications among the truly high-impact publications.	$\frac{TP}{TP+FN}$
Precision	The proportion of truly high-impact articles among high-impact articles	$\frac{TP}{TP+FP}$
F1 measurement	It considers both precision and recall and assumes that the two indicators are equally important.	$\frac{2}{F1} = \frac{1}{P} + \frac{1}{R}$
P-R curve	A dataset with N measures the generalization performance of machine learning models from the perspective of precision and recall. The curve is derived from the maximum likelihood estimation. If the P-R curve of one machine learning model completely covers the P-R curve of another machine learning model, the performance of the former is said to be better than the latter.	

N is the number of samples, $\sigma$is the estimated value before inputting the sigmoid function, and the value range of i is 1-N.

4. Implementation

The procedures for modeling the machine learning classifiers to identify PEPs in AI are depicted in Fig. 4. Discussions in this section are organized around each step except for step 4, which is discussed in Section 5.

4.1. Data acquisition and processing

4.1.1. Defining highly cited papers

A highly cited paper (HCP) is a work that "has been found useful by a relatively large number of people or in a relatively large number of experiments" (Garfield, 1979). Two approaches to defining HCPs: absolute and relative threshold (Aksnes, 2003). The absolute threshold approach defines HCPs with a fixed number, for example, 'the ten most cited papers' or 'the most highly cited papers' (Glänzel & Schubert, 1992). Aksnes (2003) and Aksnes and Sivertsen (2004) argued that the absolute threshold approach overlooks that the average citation frequency of articles varies widely across disciplines, and some leading pieces may come from highly cited fields. As a result, direct cross-disciplinary comparisons of highly cited papers are not possible with the absolute approach.

In contrast, the relative approach, such as the percentile ranks(PRs) considers an HCP (Bornmann, 2014; Bornmann et al., 2013; Bornmann & Williams, 2020) as if "it has received more than a certain multiple of the citations of the average paper within the

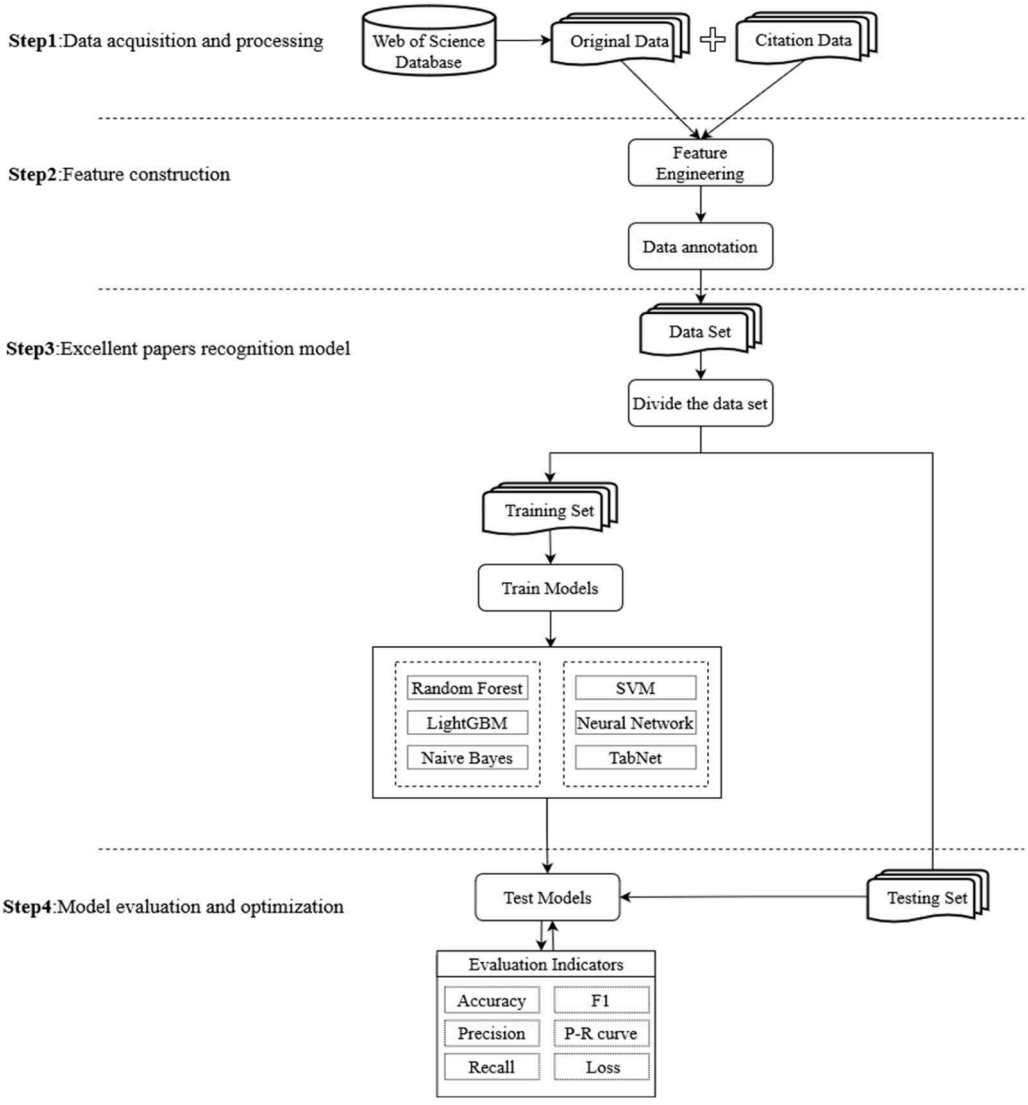

Fig. 4. Procedures for identifying potentially excellent papers.

scientific subfield" (Aksnes, 2003, p.160). A percentile is "a statistic that gives the relative standing of a numerical data point compared to all other data points in a distribution" (Lavrakas, 2008, cited in Bornmann & Williams, 2020). Research databases such as the Essential Science Indicators (ESI) of Clarivate Analytics use inverted percentiles to define HCPs. For example, ESI defines a percentile as "in which the paper ranks in its category and database year, based on total citations received. The more citations, the smaller the percentile number is. The maximum percentile value is 100, indicating 0 cites received" (http://incites.isiknowledge.com/common/help/h_glossary.html, cited in Bornmann, 2014). PRs are not affected by outliers, and their interpretation is simple (Bornmann & Williams, 2020).

In this study, we used PRs to define HCPs. According to ESI, HCPs perform in the top 1%. However, we noticed that the sample size would be too small if we used this definition. Therefore, we widened our data collection to include the top 1% and 5% of papers in the AI field between 1990 and 2010 to avoid the situation where the statistical robustness of the threshold diminishes as the number of documents declines (Tijssen et al., 2002).

4.1.2. Data

The bibliometric and citation data of 485,041 papers in AI, a sub-subject field of Computer Science, was retrieved from the Web of Science®. The publication window was set between 1990 and 2010, and the citation window was set between 1990 and 2019. This study focused on three document types: article, proceeding, and review. The papers were classified according to PRs of 1%, 5%, and 10%, non-citation, and the citation count of no more than 2 within a citation window of 10 to 30 years. Table 4 presents the distributions of papers and citations in AI from 1990 to 2010 in the original dataset. The total citations of the top 10% of HCPs in AI count for 79.00% of the total citations, which aligns with a typical long-tailed distribution. The proportions of less cited (citation frequency ≤ 2) and uncited papers are high, for example, non-citation papers alone account for 38.76% of the total number of papers.

Non-citation papers were removed. Table 5 shows the citation peaks distribution and the citation peaks intervals of 297,035 papers after removing uncited publications. Among the 297,035 cited papers, 67.62% have no citation peak, 20.48% contain one citation peak, less than 10% have two citation peaks, and only 2.46% have more than two.

4.2. Feature vector space (FVS) construction

The dataset was further screened to ensure papers in the training and test dataset have at least a citation peak. As a result, the feature vector space of 96,169 papers for machine learning was constructed by measuring and formalizing the static and time-dependent features. 96,169 papers include the top 5% HCPs and sleeping beauties with at least a citation peak. The sleeping beauties were identified by combining the K-value algorithm (Teixeira et al., 2017) and van Raan's (2004) three indicators.

4.2.1. Descriptive statistics of FVS and the AI literature

Table 6 gives examples of FVS of AI literature. In the table, d represents each publication, dn is the given publication number, and X_j represents a bibliometric feature of the papers. D_i is a document feature vector which can be expressed as $Di = (X_1, V_{i1}; X_2, V_{i2}; \ldots; X_j, V_{ij}; \ldots; X_n, V_{in})$, where $(V_{i1}, V_{i2}, \ldots, V_{in})$ is the value of the feature and $1 \leq i \leq m$, $1 \leq j \leq n$ ($m = 96{,}169$, Max, and $n = 13$, Max).

A descriptive analysis of the features (Table 7) shows that 75% of cited papers in AI have fewer than four authors, with an average of 2.89 authors. The length of papers was generally short; 75% had less than 14 pages, with an average length of 11.14. Furthermore, on average, each paper has 3.16 keywords and 24.01 references. 75% of papers were cited within the first two years of publications, and the average value of the first-citation speed is 1.64 years. The average citation count in the first year is 0.26 in the first two years is 4.44. 75% of papers were uncited in the first year, and 25% were cited once in the first two years. Per annum, citations of most papers only fluctuated a little, with an average of 1.14 times. The 75% of papers had citation peaks between 1 and 2 during the citation window time, with an average of 1.45. On average, the period from publication to first citation peak is 5.04 years and the most extended period is 27 years. The average interval between citation peaks is 6.4 years, and the most extended interval is 25 years.

4.2.2. Selection of static and time-based features of sample documents

Feature selection strongly correlates to the classifiers' performances. Fig. 5 displays the results of using Spearman to test the correlation between the features in this study. The threshold of correlation between the features was set as 0.9, and features would be removed if the correlation was greater than 0.9. The results suggested that all features are weakly correlated, and their correlations are less than 0.9. Hence, no feature was removed from the study. However, the correlation coefficient between the number of citation peaks (X_{10}) and the Interval between citation peaks (X_{11}) is −0.89, and the correlation coefficient between the time between the first

Table 4
The distribution of papers and their citations in the Artificial Intelligence field from 1990 to 2010 in the original dataset.

	Number of papers	The proportion of papers in total	Cited frequency	The proportion of cited frequency
Total	485,041	100.00%	6,467,044	100.00%
top 1%	4850	1.00%	2,665,018	41.21%
top 5%	24,435	5.00%	4,306,082	66.59%
top 10%	48,504	10.00%	5,108,749	79.00%
Cited frequency ≤ 2	289,416	59.67%	139,798	2.16%
Non-citation	188,006	38.76%	0	0.00%

Table 5
The distribution of papers and their citation peaks and the intervals between citation peaks in AI research (n = 297,035).

Number of peaks	Number of papers	Rate	Minimum peak interval	Maximum peak interval
0	200,866	67.62%		
1	60,847	20.48%		
2	28,042	9.44%	4.0	25.0
3	6290	2.12%	4.0	13.0
4	913	0.31%	4.0	8.7
5	75	0.03%	4.0	6.5
6	2	0.00%	4.6	5.2

Table 6
Examples of the feature vector space of AI literature.

	Static features					Time-based features							
Publication number	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
d2	5	1	75	9	41	1	1	9	1.62	2	7	1	1
d12	37	1	88	10	26	1	1	12	0.75	2	9	2	2
d2196	7	7	173	6	25	1	1	5	0.57	2	4	4	4
…	…	…	…	…	…	…	…	…	…	…	…	…	…
d20000	3	8	359	25	37	2	1	5	0.87	2	8	6	14
d45000	2	7	253	23	25	1	5	12	0.91	2	10	3	3
d95351	1	6	80	14	70	1	5	25	0.56	3	8	5	5
…	…	…	…	…	…	…	…	…	…	…	…	…	…
d96,169	7	14	246	18	37	1	5	38	0.35	2	9	2	2

Table 7
Descriptive statistics of static and time-dependent citation features in the sample AI literature.

Bibliometric and citation pattern features	Number of cases	Mean	SD	25%	Median	75%	Max.
Number of authors	96,169	2.89	1.50	2.00	3.00	4.00	79.00
Number of keywords	96,169	3.16	2.25	1.00	3.00	5.00	30.00
Abstract length	96,169	139.92	58.84	102.00	135.00	173.00	820.00
Article length	96,169	11.14	8.23	6.00	9.00	14.00	370.00
Number of References	96,169	24.01	19.28	12.00	20.00	30.00	913.00
First-citation speed	96,169	1.64	1.61	1.00	1.00	2.00	23.00
Citations in the first year	96,169	0.26	0.82	0.00	0.00	0.00	34.00
Citations in the first 2 years	96,169	4.44	7.05	1.00	3.00	6.00	516.00
Fluctuation of annual citation	96,169	1.14	0.59	0.74	0.99	1.38	5.39
Number of citation peaks	96,169	1.45	0.67	1.00	1.00	2.00	6.00
The interval between citation peaks	35,322	6.40	2.31	5.00	6.00	7.50	25.00
The time between the first citation peak and the publication year	96,169	5.04	3.38	3.00	4.00	7.00	27.00
The time between the highest citation peak and publication year	96,169	6.60	4.30	3.00	6.00	10.00	28.00

Note: only the papers with greater than or equal to 2 citation peaks have "the interval between citation peaks" feature, and the papers with greater than or equal to 1 citation peak have "time from first citation peak to publication year" and "time from highest citation peak to publication year" features.

citation peak and the publication year (X_{12}) and the time between the highest citation peak and publication year (X_{13}) is 0.71.

Citation peak exerts an influential role in identifying HCPs. Table 8 presents the distribution of citation peaks and the interval between citation peaks of 24,448 HCPs and 272,587 lowly cited papers in AI. Citation peak distributions between HCPs and lowly-cited papers are different. 94.08% of the HCPs had 1 to 3 citation peaks; in contrast, 73.47% of the lowly cited papers had no citation peak and remained in a lowly-cited state. As the number of citation peaks increased from 0 to 5, the proportion of lowly-cited articles rapidly decreased from 73.47% to 0. This implies that the number of citation peaks can be used to distinguish highly cited articles from lowly-cited papers, and it is a good feature to identify PEPs.

4.3. Machine learning models implementation

4.3.1. Training and test set

The dataset of 96,169 papers was divided into training and test sets according to publication period. The training set consisted of 69,418 papers from 1990 to 2007. Among them, 18,255 HCPs were marked as the positive target vector Y_i^+, whereas 51,163 lowly cited papers were marked as the negative target vector Y_i^-. The ratio between them was 1:2.8. The test set consisted of 26,751 papers, including 5592 HCPs and 21,159 lowly cited papers from 2008 to 2010, and the ratio between the two was 1:3.78.

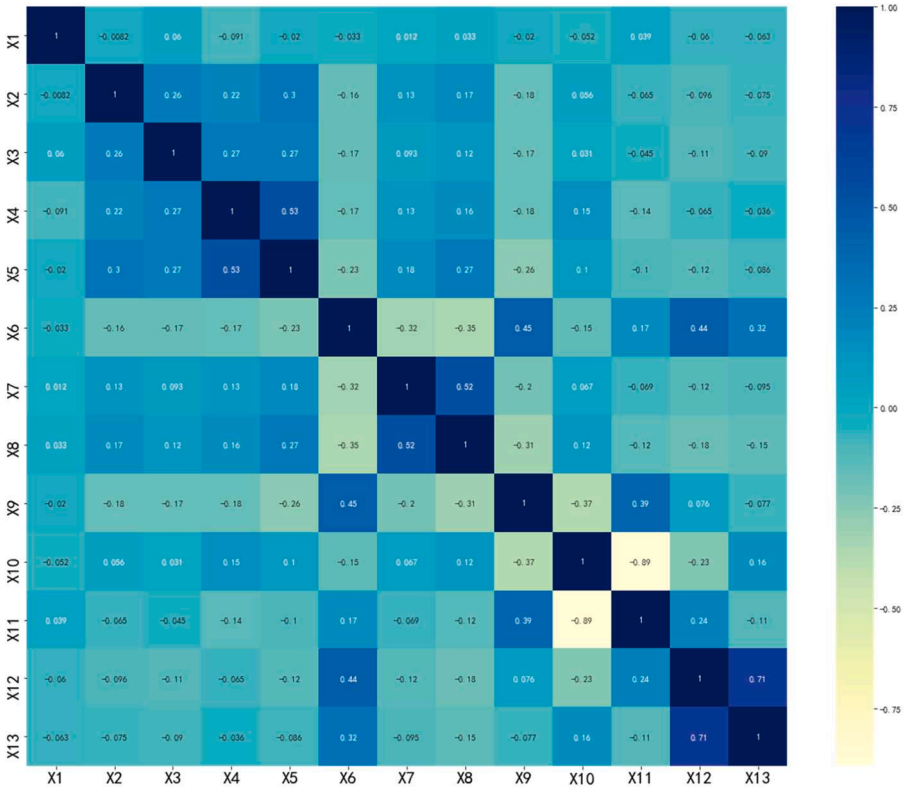

Fig. 5. Correlation coefficients between bibliometric indicators.

Table 8
Distribution of the citation peaks and intervals between citation peaks of HCPs and lowly-cited papers.

	Number of citation peaks	Number of papers	Rate%	Minimum peak interval	Maximum peak interval
Highly cited papers	0	601	2.46		
	1	8521	34.85		
	2	10,548	43.14	4.0	25.0
	3	3920	16.03	4.0	13.0
	4	788	3.22	4.0	8.7
	5	68	0.28	4.3	6.5
	6	2	0.01	4.6	5.2
Lowly cited papers	0	200,265	73.47		
	1	52,326	19.20		
	2	17,494	6.42	4.0	23.0
	3	2370	0.87	4.0	12.5
	4	125	0.05	4.0	8.7
	5	7	0.00	4.0	5.8

Table 9
Parameters for the machine learning classifiers.

Random Forest		LightGBM		Support Vector Machine		Neural Network		Tabnet	
Parameter name	Value	Parameter name	Value	Parameter name	Value	Parameter name	Value	Parameter name	Value
n_estimators	190	n_estimators	100	C	1	num_epoches	100	n_steps of decision	3
max_depth	14	max_depth	9	gamma	0.01	lr	0.01	gamma	1.3
min_samples_split	8	min_split_gain	0.5	kernel	rbf	momentum	0.8	momentum	0.02
min_samples_leaf	4	num_leaves	20			batch_size	25	Batch_size	25
max_features	4	subsample	0.8			activation function	relu	epochs	100
criterion	entropy	colsample_bytree	1			number of neurons	(3, 4)	Best epoch	98
		boosting_type	gbdt			best epoch	89	Best_val_0_mse	0.21
random_state	42	random_state	42						

4.3.2. Parameters for machine learning classifiers

Table 9 summarises the parameters obtained after optimizing the machine learning models except for the Naive Bayes because the model does not need additional parameter settings. This process combined cross-validation, manual parameter tuning and learning curve optimization of parameters to seek the optimal performance of the model. Random Forest, LightGBM, and SVM were optimized through the combination of ten-fold cross-validation and the hyperparameter learning curve. In contrast, the Neural Network model was optimized by screening the best hyperparameters and the number of neurons in multiple experiments. The TabNet was optimized through gradient descent optimization and manual tuning. As illustrated in the table, SVM was optimized as optimal when the rbf Gaussian kernel function with gamma=0.01 and L2 regularization with C = 1.

Due to the constraint of word counts, this article uses the Neural Network model as an example to demonstrate the optimization process of our EP recognition model. Fig. 6 shows the changes in the Neural Network model's recognition accuracy and binary cross-entropy loss function with increasing learning times, respectively. The horizontal axis represents the number of learning times, and the vertical axis represents the accuracy and the value of the binary cross-entropy loss function. The blue line represents the training set, and the orange line represents the test set. The model achieved the optimum training effect in the 62nd epoch during the training. In the epoch, the binary cross-entropy loss on the training set was 0.2683, and the accuracy was 0.8823. At the same time, the binary cross-entropy loss of potential excellent papers was 0.2383, and the accuracy rate reached 0.8956.

5. Results

5.1. The performance of identifying potential excellent papers

Table 10 shows the performance of six machine learning models for identifying PEPs in various indicators. All models identified PEPs with recognition accuracy between 81% and 91%. The TabNet classifier had the best performance in recognition accuracy (91%) and was followed by Neural Network, LightGBM, Random Forest, and SVM, as the accuracy rate of each was about 89%. Comparatively speaking, Naive Bayes did not perform well on this item, with an accuracy rate of 81%, and identified fewer HCPs than other models. All classifiers except for Naive Bayes had good precision rates, recall rates, and F_1 values. LightGBM and Random Forest achieved the same accuracy rate (89%) and recall rate (85%), but the former achieved better results in other indicators than the latter. The TabNet performed well in accuracy, precision, and F_1 but had a recall rate of 81% lower than LightGBM's performance. For example, the confusion matrix (Fig. 7) shows that LightGBM identified 4770 PEPs while TabNet identified 4554. Neural Network performed well in the two-class cross-entropy loss, which means that the classifier had an excellent fitting ability in the test set compared with other classifiers.

We used a confusion matrix to visualize the models' performance in identifying PEPs (Fig. 7). The row in the matrix represents the instances in an actual class, whereas the column represents the instances in a predicted class. For example, in the case of Random Forest, row 1 and column 1 shows that the model predicted 4763 truly positive papers in 5592 positive papers. The identification accuracy rate of Random Forest is 89%, and the calculation is the proportion of the correctly classified papers (TP+TN=4763+18,941) to the total number of papers (TP+TN+FP+FN=4763+18,941+2218+829).

The performances of the models were evaluated using the P-R curve. Fig. 8 shows the P-R (precision-recall rate) curve of all models except for the Neural Network and TabNet, which do not have a P-R curve. According to the average P-R scores and curves, LightGBM had the best generalization ability (0.89), followed by Random Forest (0.88) and SVM (0.86). Naive Bayes had the worst generalization ability (0.66).

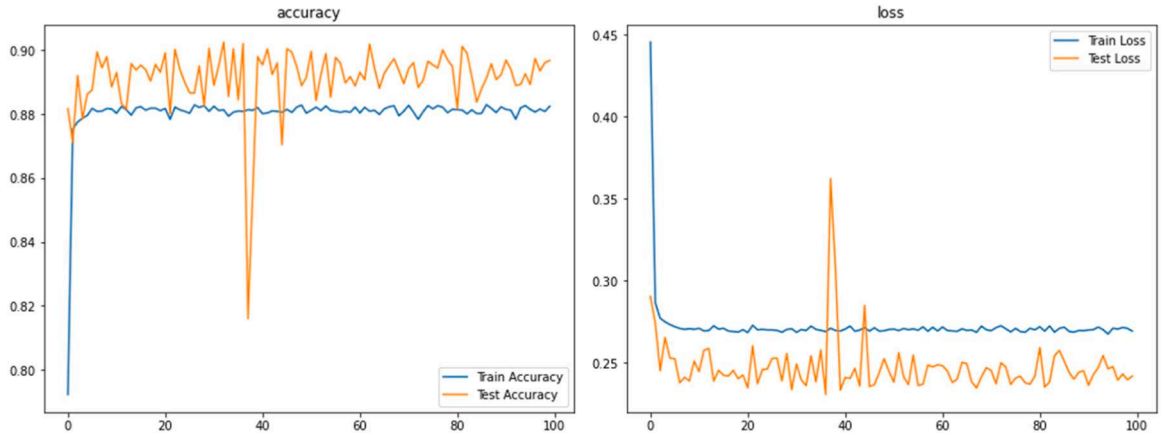

Fig. 6. The changes in the accuracy & two-category cross-entropy loss of the Neural Network classifier with different numbers of epochs.

Table 10
Performance of six classifiers of excellent papers identification.

Model Evaluation indicators	Random Forest	LightGBM	Naive Bayes	Support Vector Machine	Neural Network	TabNet
Accuracy	0.89	0.89	0.81	0.89	0.89	0.91
Precision	0.68	0.70	0.54	0.72	0.75	0.77
Recall	0.85	0.85	0.69	0.81	0.76	0.81
F_1	0.76	0.77	0.61	0.76	0.74	0.79
Average P-R score	0.88	0.89	0.66	0.86	——	——
Cross entropy loss	3.93	0.88	6.42	3.64	0.24	3.07

5.2. The significance of the time-dependent citation features in identifying PEPs

Random Forest, LightGBM, and TabNet were used to quantify the importance of the static (X_1~X_5) and time-dependent citation features (X_6~X_{13}) in identifying PEPs (Fig. 9). The results showed that the role of time-dependent citation features, especially citation peak features, in identifying PEPs is more significant than the static features. According to LightGBM, citations in the first 2 years (X_8) is the most significant feature, followed by the fluctuation of annual citations (X_9), the time between the highest citation peak and publication year (X_{13}), and the time between the first citation peak and the publication year (X_{12}). Random Forest identified the fluctuation of annual citations (X_9) as the most important feature and the first 2 years (X_8) as the next important feature. Similarly, TabNet identified the fluctuation of annual citations (X_9) as the most important feature, and the number of citation peaks (X_{10}) is the next important feature. Overall, time-based features such as X8~X13 are more important in identifying PEPs than static features X1~X5 in three models. The order of importance of features X8~X13 identified by TabNet as X9>X8>X10>X12>X13>X11 and the order by Random Forest model as X9>X8>X13>X11>X10>X12.

Given the above results, we conducted experiments on the static and time-dependent citation features separately using LightGBM. The results show that the time-dependent citation features generally performed better than the static features (Table 11). For example, LightGBM with the time-dependent citation features outperformed in accuracy (90%) and recall (71%) compared to 78% and 44% with the static features.

5.3. Characteristics of potentially excellent papers

We compared the static and time-dependent citation features of PEPs (n = 4770, Table 12) with the entire sample (n = 96,169, Table 7) and uncited papers (n = 188,006, Table 13). On average, PEPs' number of authors, keywords, and references are higher and abstract, and article lengths are longer than the entire sample and significantly higher than the uncited papers. On average of time-dependent features, PEPs' first-citation speed is 0.61 years, faster than 1.64 years of the entire sample. PEPs' slowest first-citation speed is 3 years, and the entire sample is 23. PEPs fluctuated less in annual citations, with a standard deviation of 0.12 and a mean of 0.58 compared. PEPs' mean citations of 1.02 in the first year and 14.20 in the first two years are significantly higher than the mean of 0.26 in the first year and 4.44 in the first two years of the entire sample.

Furthermore, PEPs' mean time between the first citation peak and the publication year (4.63) and between the highest citation peak and the publication year (5.49) is shorter than the entire sample (5.04 and 6.60, respectively). This implies that PEPs tend to receive more attention, and their values are likely to be realized earlier than other papers. The average number of citation peaks of PEPs (1.28) is less than the entire sample (1.45), which may be because the citation period of PEPs (2008–2010) is shorter than the entire sample's citation period (1990–2010). Comparing the intervals between citation peaks and the distribution of average citation intervals between PEPs and the entire sample with one citation peak (6 *vs.* 7.5 and 5.02 *vs.* 6.4, respectively), PEPs' citation peaks are more closely distributed. These implied that PEPs attract more attention and are cited more frequently during a specific period than other papers.

6. Robustness test of hcp threshold

The experiments have shown that the definition of HCPs and the range of HCPs used for the positive referencing vector of the training set matter to the classifiers' performance in identifying PEPs. To verify the robustness of models the study used the threshold of the top 1% to test the robustness of the models. The study repeated the same experiments using the top 1% HCP threshold. The results (Table 14) showed that all models performed better in recognition accuracy (e.g., 89% - 98%) with two HCP threshold of the top 1% and top 5%. Nevertheless, all models performed worst in recall with an HCP threshold of the top 1%, with recall rates between 59% and 76%. In contrast, the models achieved better recognition recall between 76% and 85% with an HCP threshold of the top 5% except for Naive Bayes. The test results suggest that both thresholds are valid, but because the recall rates are better with the top 5%, the study chose the top 5% to define HCPs.

7. Discussion and conclusion

The study wishes to contribute to the research on the early identification of excellent research papers. To achieve the above, the study deployed machine learning methods and selected bibliographic and citation-related features to identify PEPs. The features proposed covered two types: static and time-dependent citation features. The latter was further grouped into the citation and citation

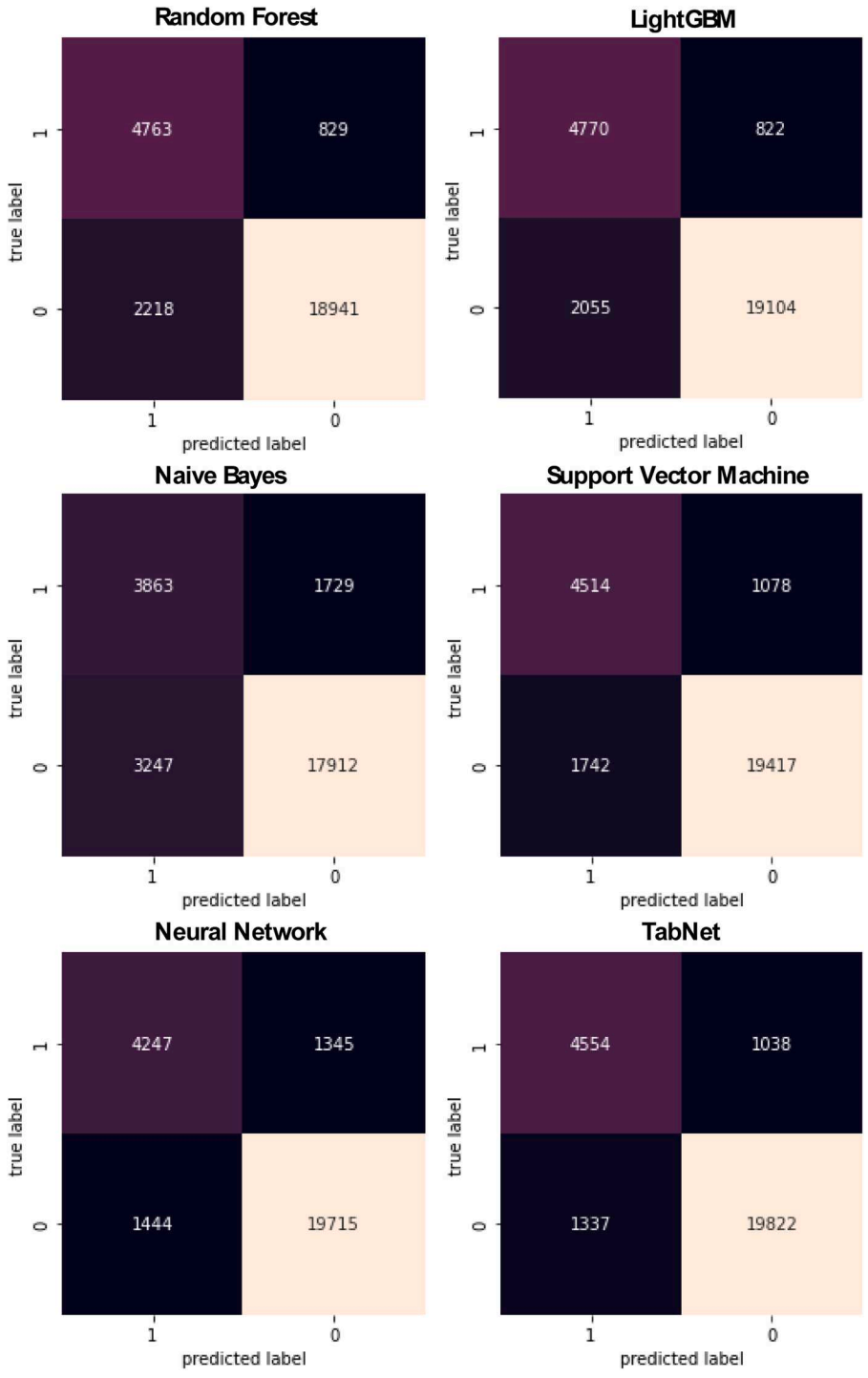

Fig. 7. Confusion matrix of 6 machine learning models' performances in recognition.

peak features. The findings suggest that the time-dependent citation features are more important than the static features, and the citation peak features are more significant than citation features in identifying PEPs. Our findings suggest that it takes PEPs in AI literature to reach their first citation peak and the highest citation peak faster than other non-PEPs (4.63 vs. 5.04). Very few PEPs have more than one citation peak (1.28 citation peak on average). The AI PEPs' average time to reach the first citation peak is faster than non-PEPs in the same field. This result is in line with the existing research that argues that a paper's influence reaches its peak within the first 5 years.

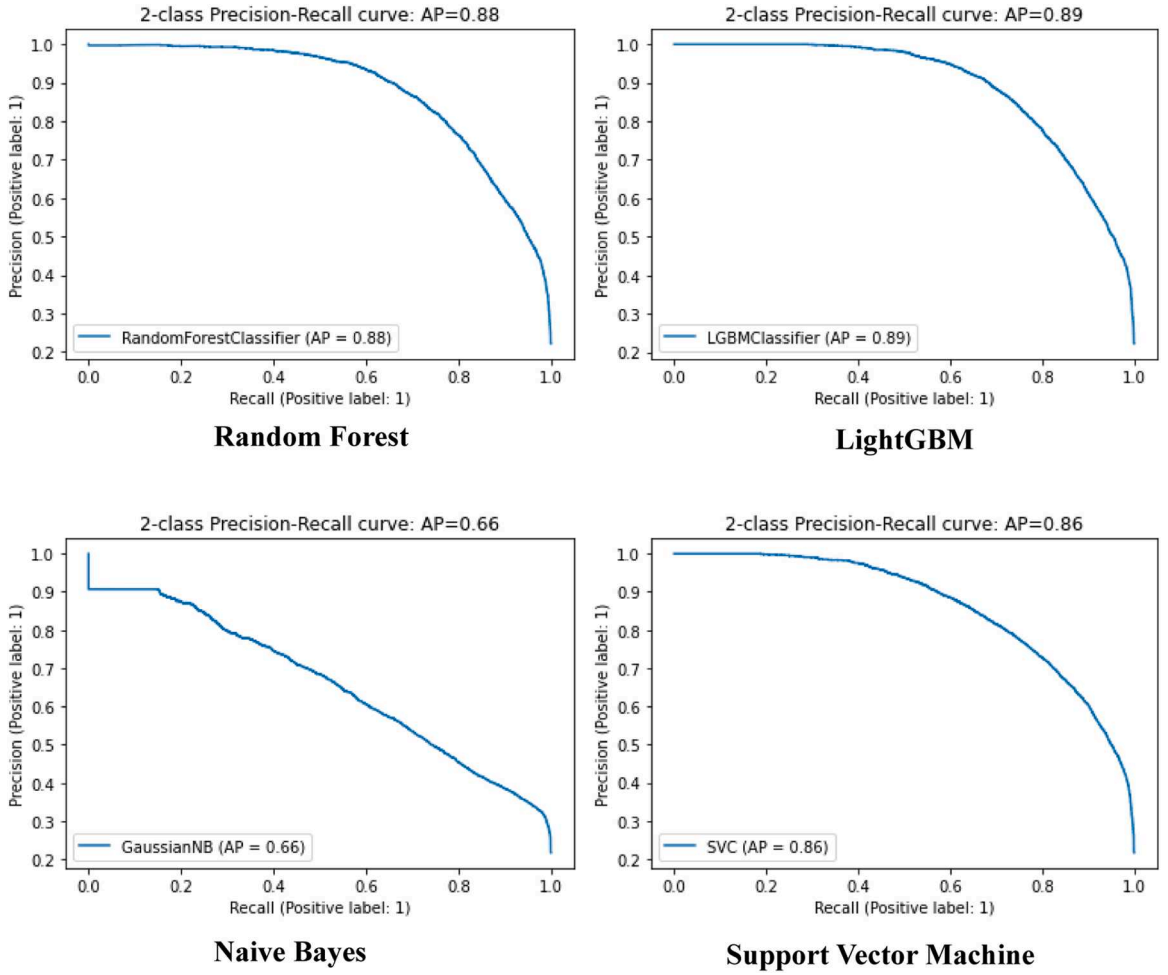

Fig. 8. P-R curve of Random Forest, LightGBM, Naive Bayes, Support Vector Machine.

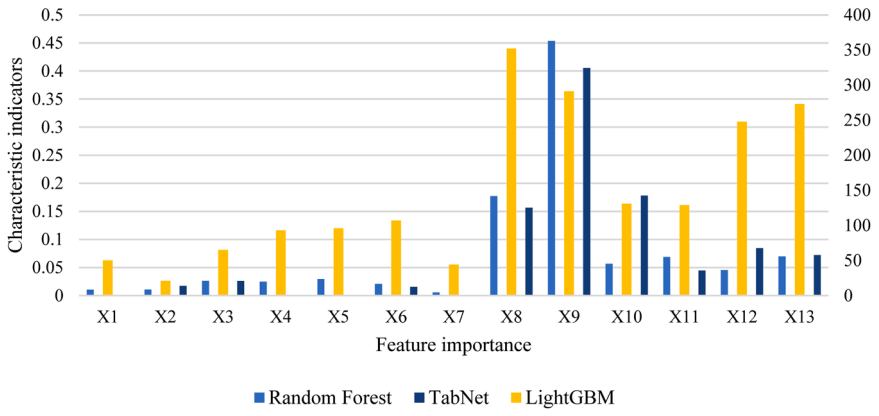

Fig. 9. Comparing the importance of the features using Random Forest, LightGBM, and TabNet.

Citation peaks can reflect the speed of knowledge diffusions and obsolete, the length of time taking a paper to be recognized, and the influence of a paper. Avramescu (1979) identified five typical citation patterns or known as citation curves, including (1) initially much-praised articles, (2) basic recognized work, (3) scarcely reflected work, (4) well-received but later erroneous qualified work, and (5) genial work (Fig. 10). These curves display different shapes of citation peaks and the aging process of a paper. The advantage of using the citation peak feature is that it can overcome the problem associated with the static features and focus on the period where the

Table 11
A comparison of LightGBM's performance in identifying PEPs based on the static features and time-dependent citation features.

Evaluation indicators	Static features	Time-dependent citation features
Accuracy	0.78	0.90
Precision	0.68	0.70
Recall	0.44	0.71
F_1	0.76	0.77
Average P-R score	0.17	0.85
Cross entropy loss	3.93	0.88

Table 12
Characteristics of potentially excellent papers ($n = 4770$).

Characteristics	Number of cases	Mean of 4770 papers	Mean of 96,169 papers	SD	25%	Median	75%	Max.
Number of authors	4770	3.14	2.89	1.68	2.00	3.00	4.00	32.00
Number of keywords	4770	4.27	3.16	2.07	3.00	4.00	5.00	25.00
Abstract length	4770	164.94	139.92	56.39	126.00	160.00	198.00	486.00
Article length	4770	13.06	11.14	8.04	8.00	11.00	15.00	84.00
Number of references	4770	37.98	24.01	25.73	23.00	33.00	46.00	379.00
The first-citation speed	4770	0.61	1.64	0.58	0.00	1.00	1.00	3.00
Citations in the first year	4770	1.02	0.26	1.82	0.00	0.00	1.00	34.00
Citations in the first two years	4770	14.20	4.44	13.58	7.00	11.00	17.00	344.00
Fluctuation of annual citation	4770	0.58	1.14	0.12	0.50	0.57	0.65	1.60
Number of peaks	4770	1.28	1.45	0.46	1.00	1.00	2.00	3.00
The interval between citation peaks	1329	5.02	6.40	1.16	4.00	5.00	6.00	9.00
The time between the first citation peak and the publication year	4770	4.63	5.04	2.19	3.00	5.00	6.00	10.00
The time between the highest citation peak and publication year	4770	5.49	6.60	2.22	4.00	6.00	7.00	10.00

Table 13
Static features of the uncited papers.

Characteristics	Number of cases	Mean	SD	25%	Median	75%	Max.
Number of authors	188,006	2.77	1.42	2.00	3.00	3.00	137.00
Number of keywords	188,006	2.54	1.86	1.00	1.00	4.00	40.00
Abstract length	188,006	107.22	58.26	75.00	106.00	140.00	1864.00
Article length	188,006	5.56	3.86	4.00	5.00	6.00	308.00
Number of references	188,006	10.10	8.32	5.00	9.00	14.00	571.00

influence of an article reaches its peak (Li et al., 2019). The initial motivation for including citation peaks in the features was based on the assumption that sleeping beauties could also be HCPs. Besides, because of the uncertainty about when the sleeping beauties will awake, using citation peak features for sleeping beauties would be more appropriate. The findings demonstrated the necessity of including the citation peak features to identify papers with delay recognition.

Research databases define HCPs with the threshold of the top 1%; instead, this study used the top 5%. The robustness test was conducted to investigate these thresholds' impact on the models' performances. The accuracy rates of using the top 1% threshold can be as high as 98% and as low as 91%, which are much higher than the top 5%. However, the recall rates of using the top 1% are poor than the top 5%. The poor recall with the top 1% threshold could be due to the limited number of HCP at the top 1%. The high recall is desirable for the output-sensitive prediction. While improving data quality can improve recall, the study demonstrates that in the case of identifying PEPs, choosing an appropriate threshold to define HCPs is equally important since the robustness of the threshold is strongly linked to the sample size (Tijssen et al., 2002).

This study identified LighGBM as the preferred model for identifying PEPs with the threshold of the top 5%. LightGBM and Random Forest achieve the same level of recall (0.85) and accuracy (0.89), but the former has slightly better precision (0.70 vs.0.68) and lower cross-entropy loss (0.88 vs. 3.93). Comparing LightGBM with Neural Network, both models achieved the same level of accuracy. LightGBM had a higher recall (0.85 vs. 0.76), although it had a lower precision rate (0.70 vs. 0.75). Its F_1 score is higher than Neural Network (0.77 vs. 0.74). TabNet, the new machine learning method, achieved good results in this study. It has the better accuracy rate of 91% and precision (0.77). TabNet outperformed LightGBM in precision (0.77 vs. 0.70) and F_1 score (0.79 vs. 0.77) but was behind LightBGM on recall (0.81 vs. 0.85) and cross-entropy loss (3.07 vs. 0.88).

The study has some limitations. First, the scope of the study is limited to identifying PEPs in the AI field. Since citation patterns vary across subject areas, it would be interesting to learn if the same results will be achieved using the same machine learning with the same features and threshold of HCPs in a different subject area. Second, the study chooses the top 5% as the threshold and provided evidence

Table 14
Robustness test using the top 1% and 5% highly cited papers as the reference standard.

	Random Forest		LightGBM		Naive Bayes		Support Vector Machine		Neural Network		TabNet	
	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%
Accuracy	0.98	0.89	0.98	0.89	0.91	0.81	0.97	0.89	0.97	0.89	0.98	0.91
Precision	0.65	0.68	0.66	0.70	0.21	0.54	0.57	0.72	0.00	0.75	0.81	0.77
Recall	0.70	0.85	0.66	0.85	0.73	0.69	0.76	0.81	0.00	0.76	0.59	0.81
F1	0.67	0.76	0.66	0.77	0.33	0.61	0.65	0.76	——	0.74	0.69	0.79
Average P-R score	0.68	0.88	0.71	0.89	0.37	0.66	0.69	0.86	——	——	——	——
Cross entropy loss	0.68	3.93	0.68	0.88	2.98	6.42	0.8237	3.64	0.08	0.24	1.09	3.07

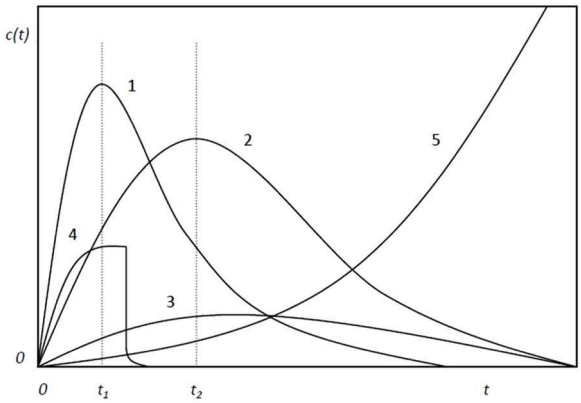

Fig. 10. Avramescu's Citation Curves of Individual Articles.

(the robustness test) to support the decision. Nevertheless, since we know the impact of threshold on machine learning models, future research can use the top 10% or else as a threshold. Third, the study assumed that all HCPs have at least one citation peak; therefore, it excluded the papers without citation peak.

Further studies should include the papers without peaks in their sample to test if the models can still identify PEPs after introducing the 'noise' (e.g., papers without citation peak) to the sample. Finally, the range of static features for this study is limited. Future studies may consider adding features from different categories (e.g., author, journal, article, reference categories) to our proposed features and investigate if different outcomes may be achieved.

Disclosure statement

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

Data availability

We have used the public dataset in our experiments.

Data availability

Data will be made available on request.

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

Abrishami, A., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics, 13*(2), 485–499.
Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research evaluation, 12*(3), 159–170.
Aksnes, D. W., & Sivertsen, G. (2004). The effect of highly cited papers on national citation indicators. *Scientometrics, 59*(2), 213–224.
Akella, A. P., Alhoori, H., Kondamudi, P. R., Freeman, C., & Zhou, H. (2021). Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics, 15*(2), Article 101128.
Abramo, G., D'Angelo, C. A., & Felici, G. (2019). Predicting publication long-term impact through a combination of early citations and journal impact factor. *Journal of Informetrics, 13*(1), 32–49.

Arik, S.Ö., & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(8), 6679–6687.
Aversa, E. S. (1985). Citation patterns of highly cited papers and their relationship to literature aging: A study of the working literature. *Scientometrics, 7*(3–6), 383–389.
Avramescu, A. (1979). Actuality and Obsolescence of Scientific Literature. *Journal of the American Society for Information Science, 30*(5), 296–303.
Bornmann, L., de Moya Anegón, F., & Leydesdorff, L. (2010). Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. *PloS one, 5*(10), e13327.
Bornmann, L. (2014). How are excellent (highly cited) papers defined in bibliometrics? A quantitative analysis of the literature. *Research Evaluation, 23*(2), 166–173.
Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of Informetrics, 7*(1), 158–165.
Bornmann, L., & Williams, R. (2020). An evaluation of percentile measures of citation impact, and a proposal for making them better. *Scientometrics, 124*(2), 1457–1478.
Didegah, F., & Thelwall, M. (2013). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology, 64*(5), 1055–1064.
Dey, R., Roy, A., Chakraborty, T., & Ghosh, S. (2017). Sleeping beauties in computer science: Characterization and early identification. *Scientometrics, 113*(3), 1645–1663.
Du, W., Li, Z., & Xie, Z. (2022). A modified LSTM network to predict the citation counts of papers. *Journal of Information Science*, 1–16.
Falagas, M. E., Zarkali, A., Karageorgopoulos, D. E., Bardakas, V., & Mavros, M. N. (2013). The impact of article length on the number of future citations: A bibliometric analysis of general medicine journals. *PloS one, 8*(2), e49476.
Fiala, D., Král, P., & Dostal, M. (2021). Are papers asking questions cited more frequently in computer science? *Computers (Basel), 10*(8), 96–107.
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.
Fu, L., & Aliferis, C. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics, 85*(1), 257–270.
Garfield, E. (1979). *Citation indexing— Its theory and application in science, technology, and humanities*. New York, USA: Wiley.
Glänzel, W., & Schubert, A. (1992). Some facts and figures on highly cited papers in the sciences, 1981-1985. *Scientometrics, 25*(3), 373–380.
Glänzel, W., & Garfield, E. (2004). The myth of delayed recognition. *The Scientist, 18*(11), 8, 8.
Hafeez, D. M., Jalal, S., & Khosa, F. (2019). Bibliometric analysis of manuscript characteristics that influence citations: A comparison of six major psychiatry journals. *Journal of Psychiatric Research, 108*, 90–94.
Hasan, S., & Breunig, R. (2021). Article length and citation outcomes. *Scientometrics, 126*(9), 7583–7608.
Huang, S., Huang, Y., Bu, Y., Lu, W., Qian, J., & Wang, D. (2022). Fine-grained citation count prediction via a transformer-based model with among-attention mechanism. *Information Processing & Management, 59*(2), Article 102799.
Iqbal, W., Qadir, J., Tyson, G., et al. (2019). A bibliometric analysis of publications in computer networking research. *Scientometrics, 119*(05), 1121–1155.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3149–3157).
Lachance, C., & Larivière, V. (2014). On the citation lifecycle of papers with delayed recognition. *Journal of Informetrics, 8*(4), 863–872.
Li, J., & Ye, F. Y. (2016). Distinguishing sleeping beauties in science. *Scientometrics, 108*(2), 821–828.
Lavrakas, P. J. (Ed.). (2008). *Encyclopaedia of survey research methods*. Thousand Oaks, CA: Sage.
Lee, S., & Choeh, J. Y. (2014). Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications, 41*(6), 3041–3046.
van Leeuwen, T. N., Moed, H. F., & Reedijk, J. (1999). Critical comments on Institute for Scientific Information impact factors: A sample of inorganic molecular chemistry journals. *Journal of Information Science, 25*(6), 489–498.
Li, L., Min, C., & Sun, J. (2019). Quantification and distribution of citation peaks. *Journal of the China Society for Scientific and Technical Information, 38*(7), 697–708.
Li, J., & Ye, F. Y. (2014). A probe into the citation patterns of high-quality and high-impact publications. *Malaysian Journal of Library & Information Science, 19*(2), 17–33.
Liang, Z., Mao, J., Lu, K., Ba, Z., & Li, G. (2021). Combining deep neural network and bibliometric indicator for emerging research topic prediction. *Information Processing & Management, 58*(5), Article 102611. Article.
van Leeuwen, T. N., Visser, M. S., Moed, H. F., Nederhof, T. O. N. J., & van Raan, A. F. J. (2003). The Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics, 57*(2), 257–280.
Van Raan, A. F. (2004). Sleeping beauties in science. *Scientometrics, 59*(3), 467–472.
Lu, W., Huang, S., Yang, J., Bu, Y., Cheng, Q., & Huang, Y. (2021). Detecting research topic trends by author-defined keyword frequency. *Information Processing & Management, 58*(4), Article 102594.
Lyu, P. H., & Wolfram, D. (2018). Do longer articles gather more citations? Article length and scholarly impact among top biomedical journals. *Proceedings of the Association for Information Science and Technology, 55*(1), 319–326.
Mistele, T., Price, T., & Hossenfelder, S. (2019). Predicting authors' citation counts and h-indices with a neural network. *Scientometrics, 120*(1), 87–104.
Ohba, N., & Nakao, K. (2012). Sleeping beauties in ophthalmology. *Scientometrics, 93*(2), 253–264.
Ponomarev, I. V., Williams, D. E., Hackett, C. J., Schnell, J. D., & Haak, L. L. (2014). Predicting highly cited papers: A method for early detection of candidate breakthroughs. *Technological Forecasting and Social Change, 81*, 49–55.
Robson, B., & Mousques, A. (2016). Can we predict citation counts of environmental modelling papers? Fourteen bibliographic and categorical variables predict less than 30% of the variability in citation counts. *Environmental Modelling and Software, 75*, 94–104.
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.
So, M., Kim, J., Choi, S., & Park, H. W. (2015). Factors affecting citation networks in science and technology: Focused on non-quality factors. *Quality & Quantity, 49*(4), 1513–1530.
Teixeira, A. A. C., Vieira, P. C., & Abreu, A. P. (2017). Sleeping beauties and their princes in innovation studies. *Scientometrics, 110*(2), 541–580.
Tijssen, R. J. W., Visser, M. S., & van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics, 54*(3), 381–397.
Tahamtan, I., & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of informetrics, 12*(1), 203–216.
Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics, 107*(3), 1195–1225.
Uddin, S., & Khan, A. (2016). The impact of author-selected keywords on citation counts. *Journal of Informetrics, 10*(4), 1166–1177.
Wang, D. S., Song, C. M., & Barabasi, A. L. (2013). Quantifying long-term scientific impact. *Science (New York, N.Y.), 342*(6154), 127–132.
Weihs, L., & Etzioni, O. (2017). Learning to predict citation-based impact measures. In *Proceedings of the 2017 ACM/IEEE joint conference on digital libraries (JCDL)* (pp. 1–10).
Wendzel, S., Lévy-Bencheton, C., & Caviglione, L. (2020). Not all areas are equal: Analysis of citations in information security research. *Scientometrics, 122*(1), 267–286.
Vanclay, J. K. (2013). Factors affecting citation rates in environmental science. *Journal of Informetrics, 7*(2), 265–271.
Wong, T. C., & Chan, A. H. (2015). A neural network-based methodology of quantifying the association between the design variables and the users' performances. *International Journal of Production Research, 53*(13), 4050–4067.
Wong, T. C., Chan, H. K., & Lacka, E. (2017). An ANN-based approach of interpreting user-generated comments from social media. *Applied Soft Computing, 52*, 1169–1180.

Xie, J., Gong, K., Li, J., Ke, Q., Kang, H., & Cheng, Y. (2019). A probe into 66 factors which are possibly associated with the number of citations an article received. *Scientometrics, 119*(3), 1429–1454.

Xu, J., Li, M., Jiang, J., Ge, B., & Cai, M. (2019). Early prediction of scientific impact based on multi-bibliographic features and convolutional neural network. *IEEE Access : practical innovations, open solutions, 7*, 92248–92258.

Yoon, S. J., Yoon, D. Y., Lee, H. J., et al. (2017). Distribution of citations received by scientific papers published in the imaging literature from 2001 to 2010: Decreasing inequality and polarization. *American Journal of Roentgenology, 209*(2), 248–254.

Yu, T., Yu, G., Li, P. Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics, 101*(2), 1233–1252.

Yuan, S., Tang, J., Zhang, Y., Wang, Y., & Xiao, T. (2018). Modeling and predicting citation count via the recurrent neural network with long short-term memory. *arXiv preprint* arXiv:1811.02129.

Zhao, S. X., & Li, J. (2015). Citation peaks in modern science: 1900–2010. *Current Science, 109*(9), 1523–1525.