Scientometrics (2023) 128:543–585
https://doi.org/10.1007/s11192-022-04547-8

A review of scientific impact prediction: tasks, features and methods

Wanjun Xia[1,2] · **Tianrui Li**[2] · **Chongshou Li**[2]

Received: 10 March 2022 / Accepted: 1 October 2022 / Published online: 26 November 2022

Abstract
With the rapid evolution of scientific research, there are a huge volume of papers published every year and the number of scholars is also growing fast. How to effectively predict the scientific impact has become an important research problem, attracting the attention of researchers in various fields, and it is of great significance in improving research efficiency and assisting in decision-making and scientific evaluation. In this paper, we propose a new framework to perform a systematical survey of scientific impact prediction research. Specifically, we take the four common academic entities into account: papers, scholars, venues and institutions. We reviewed all the prediction tasks reported in the literature in detail; the input features are divided into six groups: paper-related, author-related, venue-related, institution-related, network-related and altmetrics-related. Moreover, we classify the forecasting methods into mathematical statistics-based, traditional machine learning-based, deep learning-based and graph-based, and subdivide each category according to the characteristics. Finally, we discuss open issues and existing challenges, and provide potential research directions.

Keywords Scientific impact prediction · Citation prediction · H-index prediction · Data mining

Introduction

With the rapid development of the internet and information technology, the modes of obtaining and transmitting modes of obtaining and transmitting human knowledge have undergone major changes. The mode of obtaining knowledge has changed from traditional paper documents to electronic resources, which is more diverse, convenient and timely. In the context of the digitization of academic resources, big scholarly data, data related to different academic entities (e.g., scholars, institutions, publications and disciplines) and their relationships (e.g., collaboration and citation), have emerged (Xia et al., 2017). On the one hand, big scholarly data opens the door to the palace of knowledge for researchers.

✉ Chongshou Li
lics@swjtu.edu.cn

[1] Library, Southwest Jiaotong University, Chengdu, China

[2] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China

On the other hand, it also brings unprecedented challenges. For example, it has become increasingly challenging for researchers to quickly find influential papers from a very large amount of resources, and more information is needed to support scientific research evaluation to make research fund allocation fair (Bai et al., 2020).

Scientific impact plays an important role in the evaluation of publications, scholars, departments and institutions. The evaluation of scientific impact is usually based on past performance; however, it is more meaningful to grasp the future influence of academic entities (Cheang et al., 2014a; Fortunato et al., 2018). Therefore, predicting scientific impact is of great significance, specifically in the following aspects. Resource recommendation can help researchers quickly find papers they need and improve the efficiency of scientific research. The reliable impact prediction of researchers can help identify rising stars, facilitate expert recommendation and promote fruitful collaboration. Scientific impact prediction is also good for management of researchers and research-based institutions. Precise predictions can provide decision makers with strong evidences in many situations, such as hiring and promoting researchers, funding projects and applying for awards (Cheang et al., 2014b; Ma & Uzzi, 2018). At the same time, quantitative methods can be seen as a supplement to peer review and help allocate resources effectively. For personal development, understanding the factors affecting future academic achievements can help scholars better plan their research careers (Cheang et al., 2015; Van Dijk et al., 2014). Currently, data-driven approaches make it possible to predict scientific impact and attract the attention of researchers from various disciplines (Wang et al., 2021a; Weis & Jacobson, 2021; Xiao et al., 2021). They are based on data from digital libraries or web crawling, extracting relevant influence indicators, analyzing the laws of scientific development and realizing the prediction of the future impact of different academic entities. Related studies have been published in some prestigious journals, such as *Science* (Sinatra et al., 2016; Wang et al., 2013), *Nature* (Acuna et al., 2012) and *PNAS* (Ma & Uzzi, 2018; Way et al., 2017), indicating that this is a topic worthy of in-depth research.

Our survey mainly focused on four common academic entities, i.e., papers, scholars, venues and institutions. We retrieved data in a predefined manner from the Web of Science (WoS) Core Collection database in the period of 2000–2021. The retrieval strategy was as follows: Title = (article or paper or citation or scientific or scientist or h index or journal or institution or universit*) and (predict or forecast or long term). The document types were confined to article or review. The limitation is that papers whose title keywords are not in the retrieval strategy may be overlooked. To cover more comprehensive literature, we also selected closely related references as supplements. After filtering articles by reading their abstracts, we finally obtained 168 articles for further analysis. Among them, we identified the top 10 authors and top 5 journals that published the most studies, and they are displayed in Figs. 1 and 2, respectively. Overall, the top five journals represent 50% of all the publications. We also calculated statistics on the prediction tasks presented in each article. The results show that the number of studies predicting paper impact is the largest, followed by scholar' influence forecasting, while the prediction of institutions and venues is more complex, so the number of these articles is relatively small compared to that of papers and authors. Meanwhile, a small number of papers have attempted to predict multiple academic entities at the same time. The proportion of articles that predicting the impact of different academic entities is shown in Fig. 3.

Among these papers, there have been several review studies. Hou et al. (2019) reviewed methods and applications in prediction of paper impact, scholar impact and author collaboration. Zhang et al. (2019) summarized the author impact predictive models and the common evaluation metrics. Bai et al. (2017a) briefly introduced the methods of predicting scholarly article impact. However, these studies do not include the latest research methods, such as

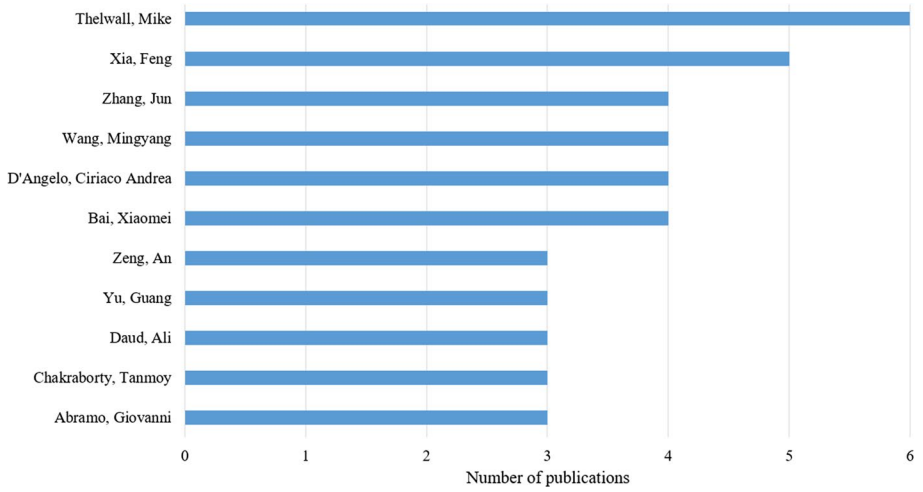

Fig. 1 The top 10 productive authors

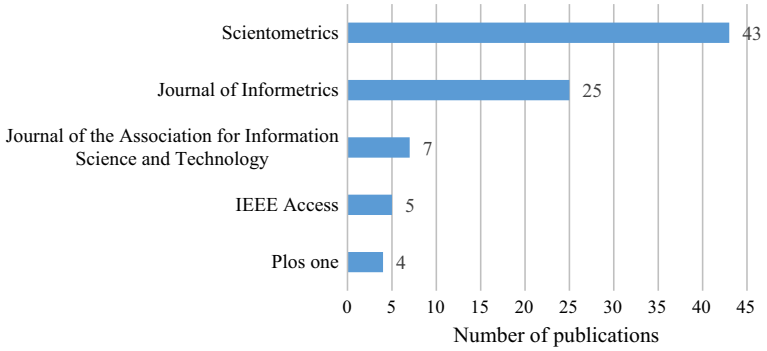

Fig. 2 The top five productive journals

deep learning and graph neural networks (GNNs), and some new features, e.g., content features and network embedding features, are not covered. In this paper, we conducted a novel and updated survey that comprehensively summarizes the prediction tasks of four types of entities (e.g., papers, scholars, venues and institutions) and the common input features and proposed a taxonomy of approaches for scientific impact prediction, involving some popular algorithms in recent years, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs) and GNNs. Figure 4 shows the framework of this survey.

The rest of our paper is arranged as follows. In Sect. "Scientific impact prediction tasks", we summarize the scientific impact prediction tasks of four academic entities and present the commonly used features. In Sect. "Taxonomy of prediction methods", different prediction methods are elaborated in detail, and datasets and evaluation metrics are also discussed. Then, the challenges and potential research directions are pointed out in Sect. "Open challenges and future research directions". Finally, we conclude the paper in the last section.

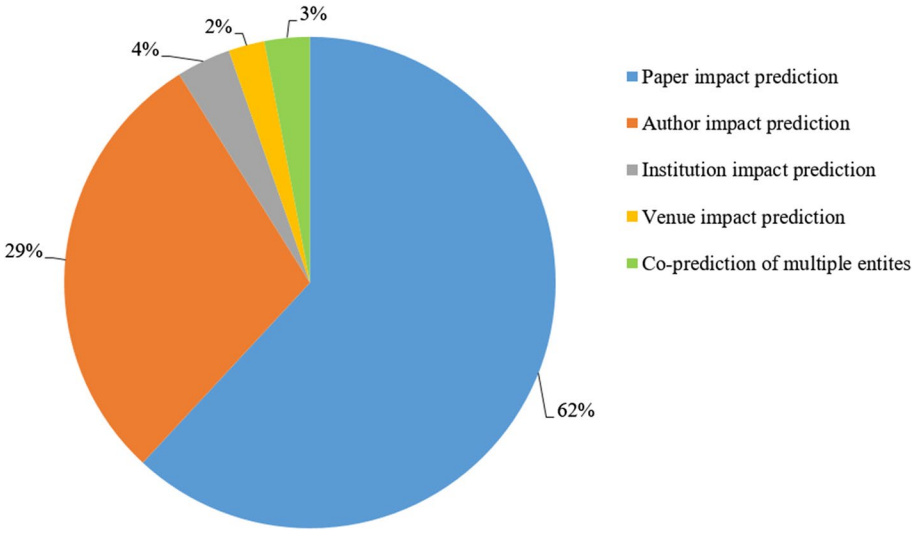

Fig. 3 Schematic diagram of the proportion of articles for scientific impact prediction of different academic entities

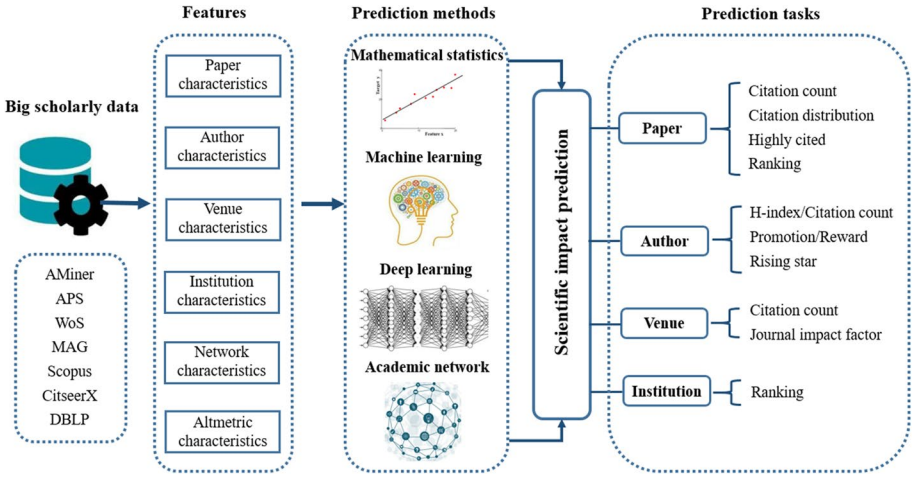

Fig. 4 The framework for this survey

Scientific impact prediction tasks

Paper impact prediction

With the large number of scientific papers that are published every year, researchers need to recognize the more influential papers in advance (Abrishami & Aliakbary, 2019). However, the rapid increase in the number of papers has also brought about information overload, preventing researchers from effectively retrieving papers and making evaluations (Zhou et al., 2021). It takes time for newly published papers to be cited, so it is valuable to

predict the citations of papers shortly after they have been published (Ruan et al., 2020). In the 2003 KDD Cup, one of the tasks was to predict the citation counts of papers. Since then, researchers have made many efforts in this field.

Prediction tasks

The goal is mainly to predict the values of the paper impact evaluation indicators. Citation count is widely used in the evaluation of paper impact, and it is simple, standard and objective, which is also the basis of many other evaluation metrics, such as the *h*-index and journal impact factor (JIF). In addition, a directed graph can be constructed from the paper citation relationship, and the evaluation of paper impact is often transformed into the importance ranking of nodes in the citation network. Therefore, there are two categories of paper impact prediction, citation-based prediction and ranking-based prediction, and they mainly include the following tasks:

(1) *Cumulative citation prediction under a given time window.* Formally, given a set of scientific publications D, the citation count of a publication $d \in D$ at time t is defined as $Cit(d,t) = \left|\left\{d' \in D : d \text{ is cited by } d' \text{ at time } t\right\}\right|$, and the goal is to estimate $Cit(t+\Delta t)$ (Pobiedina & Ichise, 2016; Yan et al., 2011). The forecast time window Δt is roughly divided into short-term (e.g., $\Delta t < 5$ year) and long-term windows (e.g., $\Delta t > 10$ year), which have no definite boundaries. The prediction task can be subdivided based on estimating the number of citations of a paper over a fixed Δt, different time intervals or several consecutive years after publication. Yu et al. (2014) predicted paper citations after 5 years of publication in the area of Information Science & Library Science, and they believed that the citation impact of the 5-year time window was an important manifestation of the quality of the paper. Yan et al. (2011) modeled the process of citation count prediction for 1 year, 5 years, and 10 years. They considered different feature combinations, and the prediction with a longer time window achieved the best accuracy ($\Delta t = 10$). Chakraborty et al. (2014) proposed a two-stage prediction model with a dataset of more than 1.5 million papers in computer science domain and Δt ranged from 1 to 5. Although these studies have claimed that their models can achieve high accuracy, there are still many challenges. There is no standard for the selection of a prediction time window, which usually depends on experience (Onodera & Yoshikane, 2015). The cumulative rate of citations in different disciplines varies greatly, so the choice of prediction time window should be different.

(2) *Long-term citation sequence prediction.* Citation time series can reflect the influence of academic publications over time (Jiang et al., 2021). The long-term citation sequence prediction task is defined as a multioutput problem, i.e., predicting future citation sequences $C_{k+1}, C_{k+2}, ..., C_n$ of the papers according to early citations $C_0, C_1, ..., C_k$ and other features. Abrishami and Aliakbary (2019) collected 175,432 papers from five prestigious journals, i.e., *Nature, Science, NEJM, Cell* and *PNAS*, and obtained their 14-year citation data. They used citation counts of a paper from the 0th to the kth year after publication as input and predicted citations of the paper from the $(k+1)$th to the nth year ($k < 7, n = 14$).

(3) *Citation distribution or trend prediction.* Because of the high uncertainty of citation prediction, it is believed that it is more useful to know the probability of citations that a publication would receive in the future. For example, Stegehuis et al. (2015) attempted to predict a future citation distribution by linking a quantile estimation technique from

extreme value theory using only the JIF and first-year citations. Instead of predicting the citation count for each paper, several studies clustered papers into several citation trends and trained a prediction model for papers with each trend, considering that papers with similar early citation dynamics may be similar in the future (Cao et al., 2016; Li et al., 2015).

(4) *Highly cited paper prediction.* Highly cited papers represent authority in the research field and have been widely used to evaluate researchers and institutions. Predicting highly cited papers in advance can help researchers track research trends and plan research directions (Wang et al., 2019a). Therefore, identifying highly cited papers is the core task of paper impact prediction. The definition of highly cited papers usually includes absolute and relative thresholds (Wang et al., 2012). For the absolute threshold, if the citation counts of a paper exceed a certain fixed value, it is considered a highly cited paper (Wang et al., 2011). However, the relative threshold is more often used, e.g., papers are ranked according to the total number of citations in a certain time window, where the top x% (e.g. x = 1,10,20) of papers are considered highly cited. Hu et al. (2020) defined the prediction task as whether the paper was in the top 25%, 33%, or 50% of total citations in the six years since publication. They extracted journal, author and keyword features in the field of marketing and management information system, and the model showed good performance for the prediction of the top 25% highly cited papers. However, their dataset covered a 4-year span, which may be conducive to early published papers. Therefore, some researchers have attempted to predict papers published in the same period that are more reasonable. Wang et al. (2019c) collected 23 features combining traditional bibliometric and alternative indicators from papers published in the same month to predict highly cited papers whose accumulated citation count reached 20% of the total citations within the dataset in a five-year window.

(5) *Early identification of sleeping beauties.* The citation patterns of scientific papers vary greatly. Normally, the citation counts of a paper will gradually reach the peak within a few years after publication and then decrease. Academia focuses more on highly cited papers. However, the value of low-cited or zero-cited papers is also worth exploring (Van Noorden, 2017). Sometimes, the importance and value of some major scientific discoveries and innovations are not recognized when they are initially published; instead, they only begin to gain attention many years later, which is referred to as delayed recognition or "sleeping beauties (SBs)" in science (Van Raan, 2004). The term "sleeping beauty" represents a special phenomenon in the scientific community that reminds us that we must have a strong sensitivity and sufficient tolerance to new ideas or discoveries. The SBs are mostly identified from retrospective research based on the time dimension (Ke et al., 2015). However, with the development of data mining, it is possible to identify SBs in advance to shorten the cycle of technological innovation and reduce the possibility of important scientific discoveries being ignored. Dey et al. (2017) developed a machine learning model to predict whether a paper is likely to become a sleeping beauty in computer science, and the results showed that SBs can be immediately identified after publication with a relatively high accuracy. To date, there have been few studies on the prediction of SBs, and further research is needed.

(6) *Paper impact ranking prediction.* Most of the abovementioned citation-based predictions require short-term historical data and are not suitable for newly published papers that have not yet begun to accumulate citations. Therefore, automatically ranking papers according to their potential impact has drawn much interest and can help researchers retrieve relevant and important information effectively (Bento et al., 2013; Zhang et al., 2018c). Different from focusing on predicting the future citations of a

paper, this task is to predict the most influential TOP *K* papers. Using the arXiv(hep-th) dataset, Sayyadi and Getoor (2009) combined information about citations, authors and publication time to calculate the future ranking score of a paper based on the citation network and paper-author network. They took the top 50 papers sorted by future PageRank as the ground truth, and there was a high correlation between the predicted score and the future PageRank score. Later, Zhou et al. (2021) proposed an age-based diffusion model with a random walk process across citation networks, which can improve the ranking of newly published papers that have zero or few citations but will be popular in the future.

(7) *Link prediction*. Citation prediction is often regarded as a kind of link prediction problem in citation networks. The goal is usually to predict the citation relationship between papers (Hou et al., 2019). Link prediction methods often use node degree to evaluate the importance of nodes. Zhou et al. (2018) used the *h*-type index and considered that nodes with high in-degree neighbors were more important to compute node similarity, which significantly improved the accuracy of link prediction in the citation network. In addition, the citation count of a paper is equal to its in-degree in the citation network; when a new link is generated, the in-degree increases, that is, the citation count increases. However, link prediction cannot capture the pattern of citation count changes over time (Liu et al., 2020). Therefore, Pobiedina and Ichise (2016) attempted to estimate the number of new links for a specific paper node by introducing a new feature named GERscore, which is based on the graph evolution rules, and the result showed that the GERscore significantly improved prediction accuracy.

Features

Researchers often pay attention to the factors that may increase the impact of their work. The citation process is complicated and is affected not only by pure scientific content but also by other factors, such as the journal in which the paper is published and the author's reputation and social influence (Tahamtan et al., 2016). The features used in paper impact forecasting include basic metadata information, content-related, network-related and altmetrics-related information.

Metadata-related information

Most of the features used in paper impact prediction come from metadata extracted from a digital library, e.g., title, abstract, keywords, references, authors and publication journal, which are relatively easy to obtain, and these features are also considered to be related to the future impact of the paper (Onodera & Yoshikane, 2015). For example, papers with more early citations are expected to have more in the future, and papers written by high-impact authors or published in high-impact journals may be more influential. In addition, researchers often perform statistical analysis on metadata information (e.g., citations or *h*-index) to obtain the maximum, minimum and average values, which are also chosen as the features of the model (Liu et al., 2020). Therefore, the basic metadata features are the most commonly used features.

Content-related information

The quality of a paper is the kernel factor that affects its readability and number of citations, but it is often ignored due to the lack of a quantitative mechanism. Singh et al. (2015) extracted two simple content-related features from the citation contexts, i.e., number of times a paper was cited within the same article and number of words within the citation context, and the results showed that these two additional features increased the prediction accuracy by 8–10%. With the development of text mining technology in recent years, some studies have begun to extract content features from metadata text to mine deeper semantic information. The topic of a paper has long been regarded as a significant feature of its content (Yan et al., 2011), e.g., hot topics or mainstream topics tend to receive more citations. Natural language processing (NLP) methods are often used to model topic diversity or keyword popularity (Chakraborty et al., 2014; Mahalakshmi et al., 2020). Hu et al. (2020) defined five keyword popularity features depending on data from Google Scholar, Google Trends and ResearchGate via a probabilistic topic model, which improved the effectiveness of highly-cited paper identification. To obtain richer semantic information, the word vector method, e.g., the word2vec and doc2vec algorithms have been used for feature extraction from title, abstract or peer review text (Li et al. 2019a; Ma et al., 2021). Moreover, some studies have paid attention to the sentiment metric of paper text. Fronzetti Colladon et al. (2020) considered semantic features of the abstract through a lexicon and rule-based sentiment analysis tool to calculate the sentiment value of each word in the abstract, and the result showed that it was better to write abstracts with more positive words. The above studies indicated that deep analysis of these content-based features can lead to further improvements in the prediction of paper impact (Ma et al., 2021; Singh et al., 2015).

Network-related information

The structural characteristics of the academic network are closely related to the node impact, and centrality is the widely used evaluation indicator. It has been found that highly cited papers have a higher betweenness centrality even at early stages after publication (Bertsimas et al., 2013). Changes in the topological position in the network can reflect the dynamics of node impact to a certain extent. Researchers have extracted various topological features to improve the prediction accuracy. Davletov et al. (2014) calculated betweenness centrality, closeness centrality, PageRank and eigenvector centrality in the citation network and built a feature vector for these metrics to predict high-impact papers. Chen (2012) provided three metrics of structural variation, i.e., modularity change rate, cluster linkage and centrality divergence to measure new boundary-spanning introduced by new paper, which can predict the future citations. In addition, inspired by word embedding technology in the field of NLP, automatically learning the vector representation of nodes in the network has become a research hotspot in recent years. Several studies adopted network representation learning approaches (e.g., node2vec and struc2vec) to capture the characteristics of citation networks, which brought additional features to the prediction of the scientific impact of papers (Luo et al., 2020).

Altmetrics-related information

New forms of academic publishing have emerged in the open scientific community, and an increasing number of papers are first published online. A series of novel measurement indicators, called altmetrics, have emerged (Thelwall & Nevill, 2018). Altmetrics refer to alternative metrics and are the creation and study of web-based metrics for analysis. They are closely related to open research activities. Data sources include article downloads and page views as well as data from social networks, news magazines, online literature management tools and public policy archives. Many publishers, e.g., *Nature, Science, Cell* and *PNAS*, use altmetrics to measure the attention of social networks to their published papers. Altmetrics are supplements to traditional bibliometrics, such as citation count and *h*-index, and they are considered a new way to measure the societal impact of research. The behavior of altmetrics occurs earlier than a citation, e.g., the download number of an article can be recorded and counted immediately, so altmetrics may be related to the future citation count of an article. Many studies used altmetrics as features for scientific impact prediction, and it has been found that these indicators are correlated with future citations to a certain extent (Akella et al., 2021; Drongstrup et al., 2020; Zoller et al., 2016).

Until now, an increasing number of features have been extracted and calculated to construct prediction models, and most studies have used multidimensional features to make predictions. As shown in Table 1, we summarize the commonly used features into six categories. However, not all of these features contribute to the final result. Ruan et al. (2020) obtained thirty features and found that only five features have significant effects on the prediction performance of the model. Therefore, correlation analysis, regression analysis and rough sets are often used for feature selection. Overall, combining two or more feature categories would result in a better prediction than using only one feature category (Hu et al., 2020).

Scholar impact prediction

The evaluation of scholars' academic performance is often based on their published papers, e.g., citations and *h*-index, which are based on past data. Therefore, if the impact of scholars can be predicted in advance, more meaningful information for decision-making, e.g., personal career development, financial support, promotion, and job offers, can be provided. Recent advances, such as data mining techniques, have made it possible to forecast the influence of scholars (Li & Tong, 2015).

Prediction tasks

Scholars' publications that are recognized or cited by their peers can reflect their academic influence. It is of great significance to understand scholars' potential or academic achievements in advance. Scholar impact prediction mainly focuses on the citation count or the number of papers published, and the prediction tasks can be divided into the following categories:

(1) *Scholar's h-index or citation count prediction.* Citation counts and the *h*-index are commonly used quantitative metrics for the evaluation of academic performance. Therefore, they have become the targets of scholars' impact prediction. Mazloumian (2012)

Table 1 The commonly used features for paper scientific impact prediction

Category	Features	Description	References
Paper	Early citation	The number of citations in the first or two years after publication	Abramo et al. (2019b); Ma et al. (2021); Ruan et al. (2020); Stegehuis et al. (2015); Stern (2014); Yu et al. (2014)
	First-cited age	Reciprocal of time interval between paper's publication year and its first citation year	Ruan et al. (2020); Wang et al. (2011); Wang et al. (2019c); Yu et al. (2014)
	Previous citation/Recent citation	Historical citations of a paper/Citation count in every year/ Citations in the past several years	Abrishami and Aliakbary (2019); Bai et al. (2019); Wang et al. (2021a); Xu et al. (2019)
	Title/Abstract length	Number of words in the title or abstract	Brizan et al. (2016); Fronzetti Colladon et al. (2020)
	Paper length	Number of pages in the paper	Bornmann et al. (2014); Haslam et al. (2008); Porwal and Devare (2020)
	Keyword count	The number of keywords in a paper or keyword repetition in abstract	Sohrabi and Iraj (2017); Xu et al. (2019)
	Reference count	The number of references in the paper	Bhat et al. (2015); Chakraborty et al. (2014); Porwal and Devare (2020); Ruan et al. (2020); Yu et al. (2014)
	Reference age/Recent reference	The average year of the references /Number of recent references since their publication	Haslam et al. (2008); Ruan et al. (2020)
	Topic Diversity	Topic distribution calculated by LDA	Chakraborty et al. (2014); Mahalakshmi et al. (2020); Wen et al. (2020); Yan et al. (2011)
	Document type	Research article or review	Liu et al. (2020); Ruan et al. (2020)
	Funding	Binary variable (yes or no) or multivariable (different levels)	Ha et al. (2016); Ruan et al. (2020)
	Paper age	Years since the paper was published	Walker et al. (2007); Zhou et al. (2020b)
Author	Productivity	The number of papers published by the first author or maximum number of papers published by the authors	Bhat et al. (2015); Danell (2011); Wang et al. (2011)
	Interdisciplinarity/Diversity	Entropy or divergence of author's publication distribution	Bhat et al. (2015); Chakraborty et al. (2014)
	h-index/Max h-index/Δh-index	h-index of the first author/Largest h-index for publication's authors/Change of h-index	Chakraborty et al. (2014); Hu et al. (2020); Mahalakshmi et al. (2020); Wang et al. (2019c); Weis and Jacobson (2021)
	Citation count/Δcitation count	Previous citation or average citation or citation increment of the first author or total authors	Danell (2011); Hu et al. (2020); Ruan et al. (2020); M. Wang et al. (2011)

Table 1 (continued)

Category	Features	Description	References
Institution	Institution number	The number of institutions in a paper	Fu and Aliferis (2010); Ruan et al. (2020); Xu et al. (2019)
	Institution prestige	Rank or quality of institution	Fu and Aliferis (2010); Wen et al. (2020)
Venue	Journal impact factor	The ratio of citations of a journal and papers published of the journal/JIF in the published year of the article or 5-year JIF	Fu and Aliferis (2010); Mahalakshmi et al. (2020); Ruan et al. (2020); Yu et al. (2014)
	Journal Rank	SCImago Journal Rank or rank based on citations	Hu et al. (2020); Liu et al. (2020)
	Citation counts	Journal total citation counts	Liu et al. (2020); Weis and Jacobson (2021); Yu et al. (2014)
	Publication number	The number of papers in the journal	Liu et al. (2020); T. Yu et al. (2014)
Social	Topological properties	Degree, closeness, betweenness, Clustering Coefficient and PageRank	Bertsimas et al. (2013); Fronzetti Colladon et al. (2020); Li et al. (2015); Park et al. (2017)
	Co-author number	Number of co-authors in a paper	Brizan et al. (2016); Li et al. (2015); Yan et al. (2011)
	Network embedding representation	Network structure calculated via node2vec or struc2vec algorithm on the citation network	Luo et al. (2020); Weis and Jacobson (2021)
Altmetrics	Online usage data	Number of views or downloads from the website	Brody et al. (2006); Ha et al. (2016); Wang et al. (2019c)
	Social media data	Number of times a paper has been mentioned, shared, bookmarked or referenced on social platforms	Akella et al. (2021); Drongstrup et al. (2020); Thelwall and Nevill (2018); Timilsina et al. (2016)

used multilevel regression models with random effects to predict future citations of a scientist's published papers, but its prediction power decayed over time. Acuna et al. (2012) considered features of the number of articles published, current *h*-index, years since publishing first article, number of distinct journals and the number of articles in high profile journals to predict the *h*-index of more than 3000 neuroscientists 5 and 10 years ahead. However, there were some restrictions, such as a career limit of 5–12 years and an *h*-index greater than 4. The validity of Acuna's equations was limited when using different datasets or considering different academic career lengths (García-Pérez, 2013). Then, some improvements, e.g., no constraints, were made. Ayaz et al. (2018) predicted the *h*-index of 15,000 scientists in computer science with different combinations of parameters, but the forecast for scholars with 1-year of work experience was inaccurate. Most of the above studies failed to distinguish scholars at different career stages, e.g., it is unfair to compare junior scholars and senior scholars together. Moreover, the *h*-index and total citation counts have cumulative advantages and are more biased toward older papers, which cannot reflect scholars' future potential. To solve these problems, Zuo and Zhao (2021) predicted future citations of future work for scholars at different career stages, which were distinguished by the number of years between the first and last publication, and they obtained more reasonable results. In addition, Dong et al. (2016) first predicted authors' *h*-indices in the next 5 years based on their previous publication records and then determined whether previously or newly published papers will contribute to the *h*-index. They also found the prediction task was more difficult for authors with high *h*-indices.

(2) *Detecting rising stars.* Rising stars often refer to scholars who currently have relatively low profiles but may emerge as prominent contributors in their field in the future (Li et al., 2009). Detecting academic rising stars can not only help scientific research institutions recruit talent but also provide candidates for reviewers, funds or award applications, which is an important task in predicting the impact of scholars. This task can be realized by ranking authors based on a potential score, classifying them into rising and nonrising stars, or clustering rising stars with similar characteristics (Panagopoulos et al., 2017). Some studies considered the degree of mutual influence of nodes in the coauthor network, which was modeled by calculating its out-link and combined with author contribution and journal ranking for iterative calculation to obtain the final score of the node (Daud et al., 2013). In addition, as the early features of high-impact scholars can provide a reference for the recognition of academic rising stars, some researchers extracted the features of scholars based on big scholarly data, and it is often formalized as a classification task to predict whether a given young scholar will be a rising star in the future. The scholar's number of publications, citation count, network indicators and journal level are chosen as features, and the citation counts of the scholar are often regarded as the classification label (Daud et al., 2015).

(3) *Scientific prize winner or promotion prediction.* Prizes and promotions are related to the recognition of scholars' academic achievements and research abilities, which also guide the direction of future scientific investments. Predicting the prize or promotion of scholars can help discover the growth characteristics and general laws of award-winning groups in social and academic activities and can provide guidance for academic evaluation and talent training. Jensen et al. (2009) studied several bibliometric indicators (e.g., *h*-index and number of papers published) to predict promotions to senior positions for CNRS researchers, but the prediction accuracy was limited. Moreover, every year, before the drawing of the Nobel Prize, institutions or individuals attempt to predict the winners of the Nobel Prize, which is also an interesting activity in the

scientific community. Citation counts have been proven to be useful in predicting Nobel laureates (Gardfield, 1977) and subsequently, Ashton and Oppenheim (1978) made improvements using nonfirst author papers, which showed better results. However, the above methods ignored the dynamic evolution of scientific development over time, and the predictive power of Nobel Prizes using simple bibliometric indicators has become limited (Gingras & Wallace, 2010). Zhou et al. (2020c) considered the prediction of Nobel Prize laureates in physics as a special binary classification task and introduced a competition mechanism considering the number of authors in the same period to normalize the citations, and the result showed that their method was effective for identifying prize winning scientists. Apart from the Nobel Prize, Rokach et al. (2011) predicted the next *AAAI* fellowship winners utilizing 292 researchers in the field of artificial intelligence. Ma and Uzzi (2018) collected more than 3000 scientific prizes, including 10,455 prize winners for over 100 years, to predict the probability that a scientist was a multiple prizewinner, and they found that prizes were more concentrated within a small group.

(4) *Scholar's publication productivity prediction.* Along with citation count and *h*-index, publication productivity is also an important indicator of scholars' academic abilities. However, scholars publish random and diverse papers; thus, it is a challenging task to predict their publication productivity (Way et al., 2017). Xie (2020) found that the number of publications within a short time interval followed a Poisson distribution and he proposed a piecewise Poisson model to predict publication productivity for researchers. However, this model is only applicable to a group of scholars and cannot be used for individual scholars. Later, he improved this method by integrating long short-term memory (LSTM) with a piecewise Poisson model, which can provide short-term prediction for individuals and long-term prediction for groups of scholars (Du et al., 2021).

Features

The prediction of scholars' influence is mainly based on their academic papers using the *h*-index and citation count. Publication-related indicators are easily accessible and quantifiable and are often used as features. In addition, scholars' personal attributes, such as age and educational background, are also related to their academic achievements. Therefore, the features of scholars' influence are summarized in our study as publication features, scholar features and social features (Table 2).

Publication features

There is a high correlation between the number of citations in the following years and the *h*-index in the previous years, and the *h*-index is better than other indicators, such as the total citation count and total paper count, in predicting future scientific achievement (Hirsch, 2007). However, several studies found different results. Schreiber (2013) found that an increase in the *h*-index was more likely to result from previous, often rather old, publications. Penner et al. (2013) thought the *h*-index was cumulative, non-decreasing indicator, which contained intrinsic autocorrelation, resulting in overestimation of its predictive power. Sinatra et al. (2016) found that the highest-impact work in a scientist's career was randomly distributed. They defined the Q parameter corresponding to the logarithm

Table 2 The commonly used features for author scientific impact prediction

Category	Features	Description	References
Publication	h-index/Δh-index	Author's current h-index or change of h-index	Acuna et al. (2012); Dong et al. (2016); Mazloumian (2012); Nezhadbiglari et al. (2016); Wu et al. (2019)
	g-index/Author Impact Factor (AIF)/Q value	Author's current g-index/AIF /Q value	de Abreu Batista-Jr et al. (2021); Ibáñez et al. (2011); Kong et al. (2020)
	Citation count	Author's total citations/Annual citations/Citation increment or average citations per paper	Ayaz et al. (2018); Ibáñez et al. (2011); Lee (2019); Mazloumian (2012)
	Number of publications	The number of the author's previous publications/ The number of publications in prestigious journals	Acuna et al. (2012); Ayaz et al. (2018); Du et al. (2021); Lindahl (2018); Xie (2020)
	Venue impact	JIF/Static ranking/Dynamic ranking	Daud et al. (2013); Li et al. (2009); Mistele et al. (2019)
	Venue citation	Citations of papers published in the venue	Dong et al. (2016); Nie et al. (2019); Zhang et al. (2017)
	Venue number	The number of venues author published	Nezhadbiglari et al. (2016); Nie et al., 2019)
Scholar (non-publication)	Academic age	Years since publishing first article	Ayaz et al. (2018); Kong et al. (2020); Wu et al. (2019)
	Gender	Male or female	Laurance et al. (2013); Van Dijk et al. (2014)
	University prestige	University rank based on the Academic Ranking of World Universities or total citations the university receives	Laurance et al. (2013); Van Dijk et al. (2014)
Social	Number of coauthors	Number of scholar's coauthors	Mistele et al. (2019); Nezhadbiglari et al. (2016); Nie et al. (2019); Zhang et al. (2017)
	Co-author prestige	Max h-index value of scholar's coauthors or the Q value of their most prolific senior co-author	de Abreu Batista-Jr et al. (2021); Kong et al. (2020)
	Co-author citation	The total or average citation of scholar's coauthor	Nie et al. (2019); Zhang et al. (2017)
	PageRank score	Author/Co-author PageRank score	Kong et al. (2020); Wu et al. (2019); Zhang et al. (2017)
	Topological properties in co-authorship network	Degree centrality, betweenness centrality, modularity and clustering coefficient	Lee (2019); Zuo and Zhao (2021)

of the number of citations in a period of time, which can predict the evolution of scientific excellence. However, cumulative indicators are not good for practical prediction (Põder, 2017), and some incremental indicators, e.g., the citation increment, *h*-index increment and incremental number of papers, that can show the dynamic changes in scholars' academic influence were extracted.

Scholar features

Individual scholars are different from each other and are affected by many factors, such as research area, age, sex, mobility and institution prestige (Yu et al., 2021). It has been found that the length of a scholar's academic career is correlated with their number of publications and citation counts and has a significant role in predicting the scientific impact of scholars (Kong et al., 2020). Researchers from prestigious institutions tend to be more productive (Van Dijk et al., 2014), and the scientific contributions of early career scholars are greatly influenced by their working institutions (Way et al., 2019). In addition, males are more likely to achieve academic success than females under equal conditions (Lindahl et al., 2020). All of these findings suggest that the nonpublication features play an important role in academic success and can be utilized for scholars' impact prediction.

Social features

Scientists' collaboration in research has become the main mode of scientific activities, which can not only promote scientific research but also help expand the influence of scholars. Junior researchers who have coauthored with top scientists can achieve a persistent competitive advantage (Li et al., 2019b). The prestige of coauthors is often seen as an important feature. Similarly, the structural characteristics of scholars in the academic network are also considered significant factors for author impact prediction, e.g., degree centrality and the average relation strength have positive effects on scholars' scientific performance (Abbasi et al., 2011).

Publication venue impact prediction

Publication venues are the carriers of academic exchange and play an important role in disseminating scientific knowledge and enlightening research ideas. Publishing articles in high-level journals is often used as an important indicator for scholars' evaluation and paper quality evaluation. Commonly used evaluation metrics, such as JIF and CiteScore, are calculated based on the publication data of the past few years, which will cause a time lag. Therefore, it is more meaningful to predict the influence of journals or to evaluate their long-term impact, which can be helpful for journal recommendations.

The journal impact forecasting task is mainly based on JIF or total citations. Wu et al. (2008) used the citations of papers to predict the JIF based on data from journals in different fields and predictions were made four months ahead of the official data. Valderrama et al. (2018) took the annual change in JIF (e.g. slope and intercept), degree of adaptation of publication guides and percentage of review articles as the independent variables to predict JIF in the field of dentistry, and the results showed a high determination coefficient. Wang et al. (2019b) considered four age characteristics of the active articles (average age, weighted average age, largest age and age of articles with largest citations) for 36 journals in the field of library and information science and found these indicators had a high

correlation with the journal's total citations, which can quantify the long-term impact of journals.

At present, the prediction of journal impact is mostly based on constructing models by extracting relevant features of published papers. However, the journal impact is affected not only by paper-related factors but also by indicators such as the review cycle, publication cycle, publication volume, publishing fees and altmetrics-related factors. To build a more accurate forecasting model, it is necessary to comprehensively consider various features in the future.

Institution impact prediction

It has become a tradition for many academic institutions, newspapers and magazines to publish rankings of research institutions or universities every year (Wilson et al., 2016), but it is still difficult to quantify the long-term impact of an institution due to the diversity of various subjective and objective factors, e.g., reputation, international collaboration, industry income and publication-related indicators. Predicting the influence of academic institutions is of great significance and can guide government agencies in making decisions, recruiting new members, guiding awards and helping students choose universities.

Considering the easy accessibility of publication data, current predictions about the impact of institutions usually shift to prediction of the institutions' publications. Related research originated from the 2016 KDD Cup competition, whose goal was to predict paper acceptance in eight top conferences in the next year, and a snapshot of the Microsoft Academic Graph (MAG) was provided for this challenge. It is believed that the prediction of paper acceptance will be helpful for the evaluation of the development potential of an institution (Bai et al., 2017b; Sandulescu & Chiru, 2016). Recently, Wang et al. (2021b) combined individual and network features to improve the ranking of the paper acceptance rate. The commonly used features are shown in Table 3. Due to the limitation of effective indicators and the complexity of the prediction task, research on institutional impact prediction is in the relatively initial stage of development.

Co-prediction of multiple academic entities

The tasks above include the prediction of an influence for a single entity. However, big scholarly data contain multiple entities and different relationships (Fig. 5). Based on these relationships between different entities, it is possible to rank the future impact of multiple types of objects in the network simultaneously. Some coranking prediction tasks are based on mutual reinforcement rules, e.g., potentially important papers published in high-quality venues and venues with good prestige attract influential researchers submitting papers. MRCoRank (Wang et al., 2016) integrates papers, authors, journals and text features into a unified framework, which can be used to predict the future influence of new publications and young researchers. Using MRCoRank, a coauthor graph, paper citation graph, venue-paper graph, venue-author graph, an author-paper graph, and paper-text feature and author-text feature graphs were built, and recent citations, recent coauthors and recently published papers were given more weight. WMR-Rank (Zhang & Wu, 2020) extracts seven types of relations to predict the future impact of papers, authors and venues through an iterative process with mutual reinforcement, and this model not only considers the time awareness but also considers the different contributions of multiple coauthors, which can be used to predict the influence of multiple entities more precisely. Moreover, Zhou et al. (2020a)

proposed a model based on a heterogeneous dynamical graph neural network to predict the cumulative impact of papers and authors to capture the dynamic processes of impact evolution and complex node interactions.

Taxonomy of prediction methods

The evolution of scientific impact is highly dynamic and complex. With the development of bibliometrics, network science and computer science, researchers have proposed many methods to solve this difficulty. Current research can be classified as mathematical statistics-based, traditional machine learning-based, deep learning-based and graph-based. The number of various methods used in annual publications is shown in Fig. 6. Early prediction methods are mainly mathematical statistics-based. In recent years, due to the rapid development of various data-driven models, machine learning, deep learning and graph-based methods have been used widely.

Mathematical statistics-based methods

Statistical learning is the most widely used method in scientific impact prediction. Statistical learning can analyze various features of academic entities, and can establish mathematical models to find out the relationship between relevant features and scientific impact to fulfill the prediction tasks, such as citation count prediction and JIF prediction (Fig. 7). The influential features are identified from a set of candidate variables to build the prediction model, and stepwise regression (Yu et al., 2014), negative binomial regression (Onodera & Yoshikane, 2015), ordinary least squares regression (Abramo et al., 2019a), quantile regression (Danell, 2011), hierarchical regression (Ha et al., 2016) and semi-continuous regression (Klimek et al., 2016) are commonly used algorithms and the citation counts are used to generate the equation, but this is not considered strictly prediction due to the possibility of system changes between years (Thelwall & Nevill, 2018). Statistical learning selects important independent variables through feature selection methods, but there is no consensus on the choice of independent variables. The main reason is that the existing models assume that multiple factors are independent of each other and do not consider the interaction between them. In addition, the sample of the multiple regression model is limited to a specific field and the generality of the conclusion may be limited.

In addition, a mathematical statistics model can describe the process of citation accumulation. Wang et al. (2013) derived a model to forecast long-term citations of paper based on three parameters: preferential attachment, aging and fitness, and the citation dynamics of paper i at time t can be described as:

$$c_i^t = m\left(e^{\lambda_i}\Phi(\frac{\ln t - \mu_i}{\sigma_i}) - 1\right) \tag{1}$$

where the parameter set $(\lambda_i, \mu_i, \sigma_i)$ can be calculated based on its historical citation, when $t \to \infty, \Phi \to 1$, the ultimate impact, which represents the total citations a paper acquired during its lifetime, can be obtained by: $c_i^\infty = m(\lambda^i - 1)$. It means that the ultimate impact of a paper is only related to the relative fitness λ. Although this method can model citation dynamics, it performs poorly in some disciplines and is prone to overfitting (Cao et al.,

Table 3 The commonly used features for institution scientific impact prediction

Category	Features	Description	References
Institution	Research ability	Rank or historical relevance scores of each institution, calculated by the number of papers published	Bai et al. (2017b); Qian et al. (2016); Sandulescu and Chiru (2016)
	Geographic feature	Geographic distance between an institution with the conference location	Bai et al. (2017b)
	Economic feature	State GDP where the institution is located	Bai et al. (2017b)
Author	Author impact	AIF, Q-value or h-index and their statistics feature (standard deviation, sum, minimum, maximum, median and mean)	Bai et al. (2017b); Sandulescu and Chiru (2016)
	Activity	Measured by the number of papers published in a certain period of time	Wang et al. (2021b); Wilson et al. (2016)
	Continuity	Number of authors published in consecutive years in the same conference	Shuang (2016)
	Productivity	Number of first authors/Number of authors who published at least one paper	Qian et al. (2016)
Network	Institution collaboration network	Degree centrality, betweenness centrality, PageRank score	Shuang (2016); Wang et al. (2021b)
	Co-author network	Author collaboration score, measured by collaboration frequency between authors	Wang et al. (2021b)

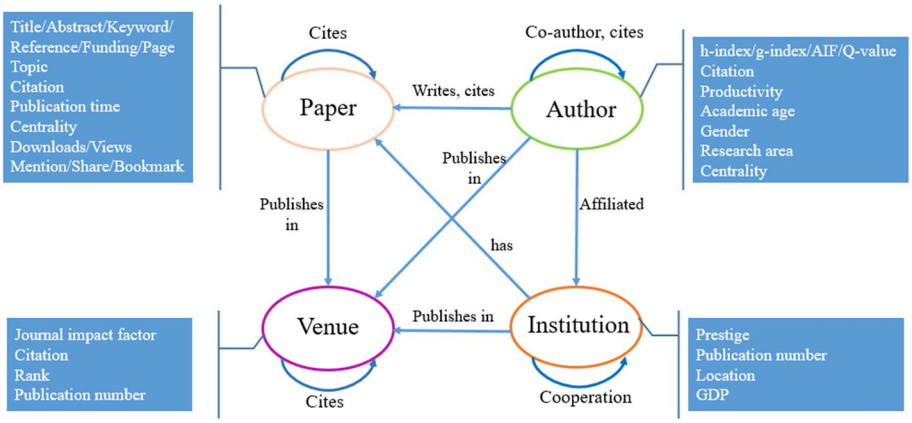

Fig. 5 Entities and their relationships in academic network

2016), which has been improved by later studies (Bai et al., 2019; Shen et al., 2014). In addition, Sinatra et al. (2016) formulated a stochastic model that can accurately predict the evolution of a scholar's impact. They defined a unique parameter Q for each scientist, which was formulated as:

$$Q_i = e^{\log c_{10,i} - \mu p} \tag{2}$$

where $c_{10,i}$ is the average citation of papers published by scholar i in recent 10 years, p is the potential impact of the research topic. Q-value is highly stable and can compare scientists of different ages and stages, but it requires a long period of observation and is not suitable for new scholars.

Traditional machine learning-based methods

With the development of big data and large-scale computing in the past ten years, machine learning has performed well in many prediction tasks. In the context of big scholarly data, traditional machine learning methods are widely used in scientific impact prediction and obtain high accuracy.

Supervised learning

Supervised learning is a category of machine learning methods in which models are trained using labeled data. The training set includes input features and output variables, and the goal is to learn the mapping from input to output to make predictions about unknown data. Some studies on the prediction of paper citations and scholars' h-indices have used supervised learning methods, and indicators such as paper-related and author-related indicators are often chosen as input features (Fig. 8). Support vector regression (SVR), random forest (RF), K-nearest neighbor (KNN), linear regression (LR), BP neural network and gradient boosting algorithms are often used to train the model (Livne et al., 2013; Ruan et al., 2020; Singh et al., 2015). Ruan et al. (2020) collected literature

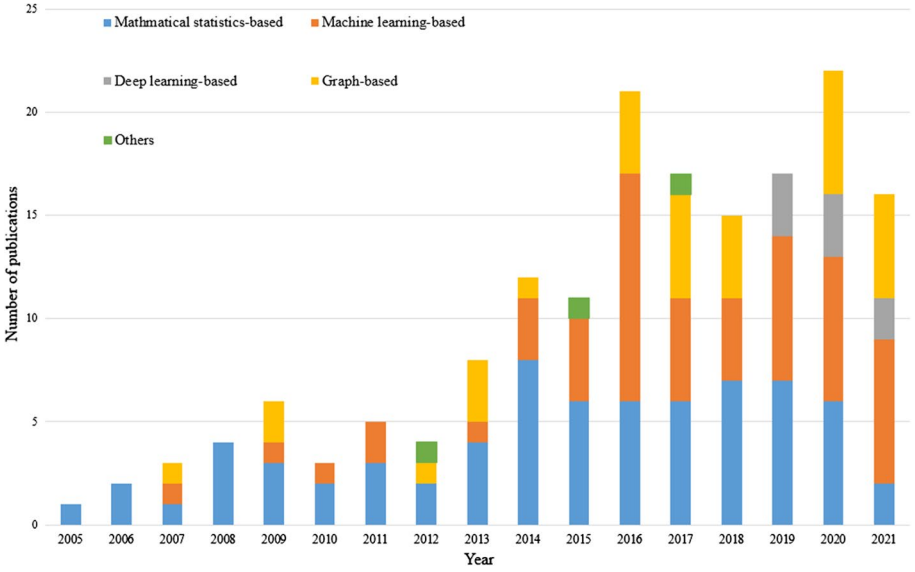

Fig. 6 The number of different methods used in annual publications

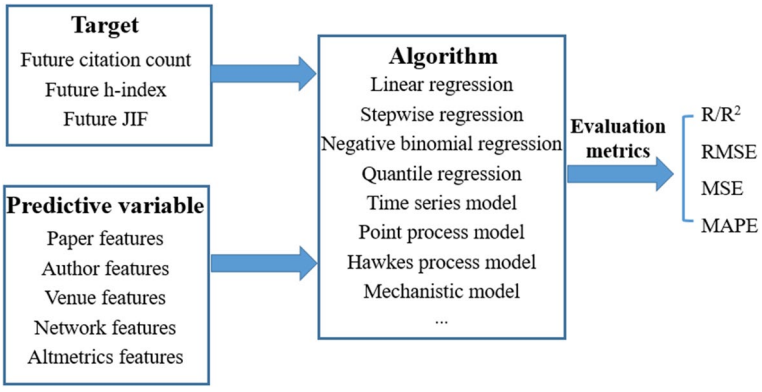

Fig. 7 Schematic diagram of mathematical statistics based-methods

published from 2000 to 2013 and citation data before 2018 and applied the four-layer BP neural network to predict 5-year citations of nearly 50,000 papers. They chose the Adam optimizer and ReLU activation function to train the model, and L2 regularization was used to prevent overfitting. The mean squared error (MSE) and R^2 for the test datasets showed that the BP neural network outperformed other algorithms, such as SVR, KNN, RF and XGBoost. Weihs and Etzioni (2017) used RF and gradient boosted regression trees (GBRT) with more than 20 features to predict the author *h*-index with a dataset of four million computer science papers written by approximately 800,000 authors. Compared with Acuna's model (Acuna et al., 2012), the best prediction accuracy rates of 5 and 10 years increased by 24.8% and 50.6% respectively. Mistele et al. (2019) trained a

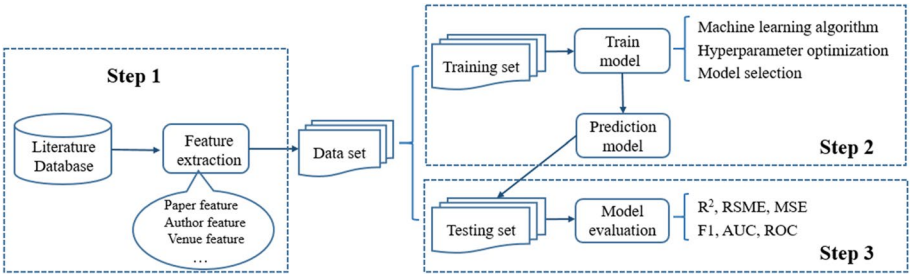

Fig. 8 The process of machine learning methods in scientific impact prediction

feedforward neural network to predict authors' citations and *h*-indices based on arXiv publications in physics field, and they claimed the R^2 of 10-year prediction was higher than Weihs and Etzioni's method, but they used different datasets.

The citation counts of papers and the *h*-indices of authors follow a heavy-tailed distribution, which may skew the prediction (Dong et al., 2016). Error metrics for continuous loss functions are difficult to interpret (Fu & Aliferis, 2010). Instead of predicting the actual future citations, many researchers regard scientific impact prediction as a classification problem. Generally, recognizing highly cited papers is often defined as binary classification, and different classification labels are set in advance. Fu and Aliferis (2010) used an SVM with a heterogeneous polynomial kernel to develop a binary classification model to predict whether an article would exceed T citations (T = 20,50,100 and 500) within 10 years, and the AUC ranged from 0.86 to 0.92. Wang et al. (2019c) collected more than twenty features and utilized three feature selection techniques to reduce redundant features. They defined three classes, i.e., whether the citations after 3 years were highly-cited, medium-cited or low cited, and employed naive Bayesian (NB), KNN and RF on the obtained features, the experiment showed all the average classification accuracies were above 0.9, but their dataset was relatively small with only 617 articles. Bhat et al. (2015) defined 2-class (whether the paper was zero cited or not) and 3-classes (0, 33-rd and 66-th percentiles of the citation distribution), and applied NB, SVM, RF, boosted trees to a large dataset with over 800,000 papers, and the best classifier (RF) yielded accuracy of 0.87.In addition, Nie et al. (2019) formalized a binary classification with five different classifiers, i.e., KNN, RF, SVM, GBDT, XGBoost to predict whether the given young scholar would be a rising star in the future, and the label information was based on the increment of impact score calculated by the quality of citing papers and the influence of coauthors, and the best performance achieved F1 score close to 0.8.

Unsupervised learning

Unsupervised learning is a kind of machine learning where the data are unlabeled, and the dataset can be classified according to the similarity between samples. Paper citations are highly random and the evolution and dynamics of a scholar throughout the scholar's whole academic career are also different. Therefore, it is difficult to characterize the dynamics of scientific impact. Inspired by the above problems, some studies divided paper citations or scholars into different types and then predicted the impact of different groups. Cao et al. (2016) found L previously published papers that matched the citation dynamics of the test

paper with the smallest matching error by calculating the Euclidean distance and then clustered them into K clusters by fitting a Gaussian mixture model, which can obtain K possible trends of the paper's future citations and probabilities. Panagopoulos et al. (2017) used the evolution of author features (i.e., productivity, impact and collaborative indicators) over time as the input to K-means, which clustered the authors into seven categories and the "rising stars" cluster can be detected through the biggest improvement over time across all of the key performance indicators.

Deep learning based-methods

Apart from traditional machine learning methods, in recent years, deep learning has shown outstanding performance in various research fields, such as computer vision and NLP. Deep learning is not only an effective method for the prediction but also a powerful tool for feature extraction. Studies have shown that feature design based on deep learning effectively improves the feature extraction of academic entities in terms of metadata text and network structure (Ma et al., 2021).

RNN and its variants

The citations of papers have time characteristics, and it is more accurate to save the information in the time series. Researchers have attempted to learn predictive models based on citation sequence patterns with early information through deep learning structures such as RNNs, LSTMs or gated recurrent units (GRUs) (Abrishami & Aliakbary, 2019; Wen et al., 2020; Yuan et al., 2018) that are mainly suitable for long-term citation prediction. Yuan et al. (2018) proposed a many-to-one model with two-layer LSTM units and integrated four major phenomena, i.e., the intrinsic quality, as represented by the citation count serving as the input of the model, the aging effect and the Matthew effect, which can be modeled by the forget gate and update gate, respectively, and the recency effect, which can represent current working memory. They used the citation data five years after the paper was published as training data to predict the citations in the next five years, and their model outperformed traditional machine learning algorithms such as SVR and LR. Abrishami and Aliakbary (2019) designed a many-to-many RNN architecture only with the early citations as the input sequence. They made experiments with different input sequence lengths ($0 < k < 7$) and the results showed that the model can perform better than other baselines when the input sequence $k >= 3$, but they ignored many other time series features. Ma et al. (2021) used the doc2vec algorithm to encode metadata text and developed a Bi-LSTM model with an attention mechanism to extract paragraph-level semantic information, from which an early citation vector was combined as input, and then predictions for the next 8 years were realized through two fully connected layers. The prediction accuracy was higher than that of GBRT and XGBoost. The above models also found that the prediction performance will improve with the length of early citation sequence increasing, which may due to the fact that longer citation history can include more information, but it can also affect the timeliness of prediction in some research field (e.g., computer science), especially those newly published papers.

CNN-related methods

CNNs have been used to capture complex temporal patterns of citations, and they can automatically transform the feature from the initial representation to a higher-level representation and learn the mapping from input to output. Xu et al. (2019) designed a CNN model with three convolution layers to capture the complex nonlinear relationships between the early network features and the final citation count to predict the long-term citation count of papers in the field of Markov chain, and the results showed that the R^2 of 5-year prediction can reach 0.9134, exceeding the comparison model by 5%. Wang et al. (2020) constructed an attention CNN model to predict paper citations. They used the doc2vec and word2vec algorithm to vectorize the paper text, and merged with journals and altmetric features to build feature matrix, then an attention layer was added to focus on key features, and the model had a higher accuracy than LR and classification and regression tree (CART).

Graph-based methods

Graph-based methods are likely more effective because they consider information from the network structure. In an academic network, the papers, authors and venues are treated as nodes, and citation relationships and coauthor relationships are treated as edges. The network structure can be divided into homogeneous and heterogeneous structures, and different network structures are chosen according to different prediction tasks. The commonly used network types include citation networks, coauthor networks, author-paper networks, and paper-journal networks. Therefore, the first step is to construct a suitable academic network and then to realize the scientific impact prediction by mining the hidden relationships in the graph. Figure 9 shows the tasks and the corresponding network structures.

PageRank-like algorithms

Inspired by the web page ranking problem, PageRank-like algorithms are widely used in networks to rank academic entities. The more citations the paper has, the higher its PageRank value and the greater its influence. If a paper with a higher PageRank value cites other papers, then the PageRank of the cited paper will be higher. However, the citation network is special in that it is a temporal network, and PageRank is biased toward older papers, making it difficult to predict newly published papers without many citations. For the prediction of paper influence ranking, researchers have considered the time decay mechanism (e.g., exponential decay) and integrated it into the diffusive random walk process, which assigns more weight to recent papers, and the node scores can predict the future popularity of papers (Walker et al., 2007). Researchers considered the real process on citation networks based on the probability that a researcher will follow a paper's references decays with the increase in each diffusion step, and the results showed that their model can predict newly published papers that may become popular in the future well (Zhou et al., 2020b). In addition, the ranking of a scholar's future impact, especially the discovery of rising stars, is often modeled by the mutual influence of scholars in academic networks. Using PubRank (Li et al., 2009), a coauthor network, which was weighted and undirected, was constructed, and the weights of edges were mutually

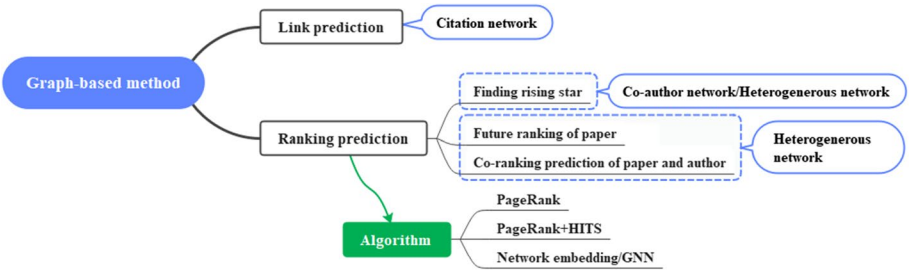

Fig. 9 Application of the graph-based methods

influenced, as calculated by the number of publications coauthored, and node weights were assigned by the quality of a researcher's publications, which was formalized as:

$$PubRank(p_i) = \frac{1-d}{N} + d \cdot \sum_{j=1}^{|V|} \frac{\omega(p_i, p_j) \cdot \lambda(p_i) \cdot PubRank(p_j)}{\sum_{k=1}^{|V|} \omega(p_k, p_j) \cdot \lambda(p_k)} \tag{3}$$

where $\omega(p_i, p_j)$ is mutual influence, defined as the number of coauthor papers between p_i and p_j divided by the number of papers of p_j. $\lambda(p_k)$ is publication quality score. A series of PubRank scores over several years for each author was calculated and scholars with larger PubRank score gradients were identified as rising stars.

Based on PubRank, a variety of improved methods that considered the author order, coauthor citation and dynamic ranking of venues were proposed for calculating node and edge weight (Daud et al., 2013, 2017). PageRank-like algorithms are suitable for homogeneous networks that contain only one type of node and link relationship, e.g., paper citation networks or coauthor networks. However, the structure of a homogeneous network is simple, and it is possible to ignore important information.

PageRank + HITS algorithms

The academic network structure is highly complex and heterogeneous, and researchers have divided the heterogeneous academic network into multiple subnetworks based on the relationships between authors, papers, venues and institutions. For subnetworks of a single node type, PageRank is used to calculate the importance of the node, while for bipartite graphs, HITS is often used for calculation, and each subnetwork is calculated independently. Finally, the scores of each node are weighted and fused. FutureRank (Sayyadi & Getoor, 2009) was the first algorithm used to rank the future impact of papers; PageRank and HITS were used to integrate the paper's PageRank value, author's authority value and time weight in the iterative process. The formulation is shown as:

$$R^p = \alpha \cdot M^C \cdot R^C + \beta \cdot M^{A^T} \cdot R^A + \gamma \cdot R^{Time} + (1-\alpha-\beta-\gamma) \cdot \frac{1}{n} \tag{4}$$

where M^C is the citation matrix, M^A is the authorship matrix, $M^C \cdot R^C$ is the PageRank score in the citation network, and $M^{A^T} \cdot R^A$ is the authority score in the authorship network. $R^{Time} = e^{-\rho t}$ gives more favor to recently published papers.

MRFRank (Wang et al., 2014) was used to improve the time weight design method, taking the cited time into account and adding text features into the network, which significantly improved the performance. However, these methods manually design time-aware weights that cannot model the dynamics of academic networks well. To solve this problem, a heterogeneous scientific hyper network framework (HSHMRR), consisting of seven subnetworks, was defined and combined with the learning-to-rank algorithm multiple additive regression tree (MART), which can capture the dynamic nature of academic networks (Zhang et al., 2018c). The experiments based on MAG showed that the HSHMRR-MART outperformed FutureRank by 24%-29%. For author impact prediction, CocaRank (Zhang et al., 2016b) and ScholarRank (Zhang et al., 2016a) divided heterogeneous networks into paper citation network, paper-author network and paper-journal network, which can contain more information and make the prediction of a rising star more reasonable.

The above graph-based method directly used the weights of edges to construct the relationship matrix between entities, retaining the global structure information of the network, but they ignored the importance of the local structure information of the network to the evaluation of the influence of the nodes in the academic network. In addition, different networks were simply merged into the random walk framework, and they cannot effectively learn heterogeneous network structure information and other information. Meanwhile, existing graph-based methods heavily depend on the global structure, ignoring the local structure information, which restricts the prediction accuracy.

Network embedding and GNN

PageRank-based methods require a large number of calculations, and performing these calculations takes considerable time when the number of nodes is large. In recent years, network embedding, which aims at learning the low-dimensional latent representation of nodes in a network, can preserve local and global information and can improve efficiency. Xiao et al. (2019) provided a network embedding model that can take the global, local structural and text into account simultaneously. They constructed a paper citation network with text information, a coauthor network and a paper-author network. They used KL divergence to describe the difference between the probability distribution of each node in the latent vector space and that of the node in the network. The subnetworks were combined to minimize the objective function, realizing the representation learning of the nodes, and the future impact of papers and authors were mutually ranked by integrating the learned embedding representations into a multivariate random-walk process. This model made full use of text information to help learn the potential similarities between papers and can learn better vector representations of newly published papers that lacking citations. In addition, the academic network is dynamically evolving, with new papers, new authors and new links generated every year; therefore, it is important to rank their future influence in a dynamic graph. Generally, it is believed that there is a relationship between a paper's citation and that of its neighbors (Holm et al., 2020). GNNs are suitable for structured prediction problems due to the neighborhood changes caused by the graph topology (Cummings & Nassar, 2020). Jiang et al. (2021) utilized relational-GCN which was extended with a simple temporal alignment technique to learn the embedding of metadata nodes in a dynamic heterogeneous information network, and this model can predict a new paper's citation time series without leading citations.

Model comparison and analysis

Comparison of different methods

Table 4 shows advantages and disadvantages of different methods. Mathematical statistics models construct rigorous mathematical formulas, explain causality with the support of mathematical theories and discover the changing laws of scientific influence. Machine learning methods make full use of high-dimensional features and nonlinear relationships in scholarly big data to build models and obtain high prediction accuracy, while graph-based methods can utilize available structural information, such as the citation network and the author network to enrich features and model the dynamic changes of academic entities. Different methods have their advantages and application scenario, but with the opening and accumulation of data, machine learning-based, especially deep learning-based methods, and graph-based methods will play a greater role in academic data mining and entity relationship mining.

Dataset

The datasets used for scientific impact prediction usually come from the digital libraries or academic search engines. Figure 10 shows the sources of datasets commonly used in the surveyed articles.

These datasets are mainly divided into three types: (i) Subscription access. Digital libraries play an important role in the storage and acquisition of global academic information. To complete the prediction task, researchers often collect data based on a certain subject area or different journals from the WoS or Scopus database. These databases are multidisciplinary that can meet the needs of researchers in multiple fields, but they need to be subscribed and the data sets are often not publicly available. (ii) Free access. With the large-scale digitization and explosive growth of academic resources, free-to-use academic search engines have emerged, which can help researchers access online resources more easily. Therefore, researchers often extract data from these academic search engines such as Google Scholar, MAG, Semantic Scholar, DBLP, etc. However, they also need to retrieve data and build datasets manually, and in terms of complex feature extraction, e.g. citation relationship, this is often time-consuming. (iii) Public dataset. To better promote scientific progress, researchers or some data mining competitions have released the datasets used and the common public data sets are shown below.

AMiner[1] It is an academic data analysis and mining platform developed by Tsinghua University, containing more than 200 million articles. Some datasets have been published, including research data, such as citation network analysis, expert discovery and name disambiguation.

APS[2] It is comprised over 450,000 articles published in American Physical Society journals since 1893. It contains two data files: article metadata and citing article pairs.

[1] 1 http://www.arnetminer.org/data.
[2] 2 https://journals.aps.org/datasets.

Table 4 Advantage and disadvantage of different methods

Method	Advantage	Disadvantage	Evaluation metric
Mathematical statistics- based	Interpretable	Limited prediction accuracy Cannot handle high-dimensional features	R^2, RMSE, MAE, MAPE
Machine learning-based	High accuracy	Lack of interpretability Need parameter tuning Feature engineering is time-consuming	Accuracy, F1, AUC, MSE, MAPE
Deep learning-based	Automatically extract hidden features and have high accuracy	Lack of interpretability Need to learn a lot of parameters	Accuracy, Recall, F1, AUC, MSE, MAPE
Graph-based	Contain structural information	High computational complexity Cold start	TOP k ranking accuracy, NDCG@k, SPCC

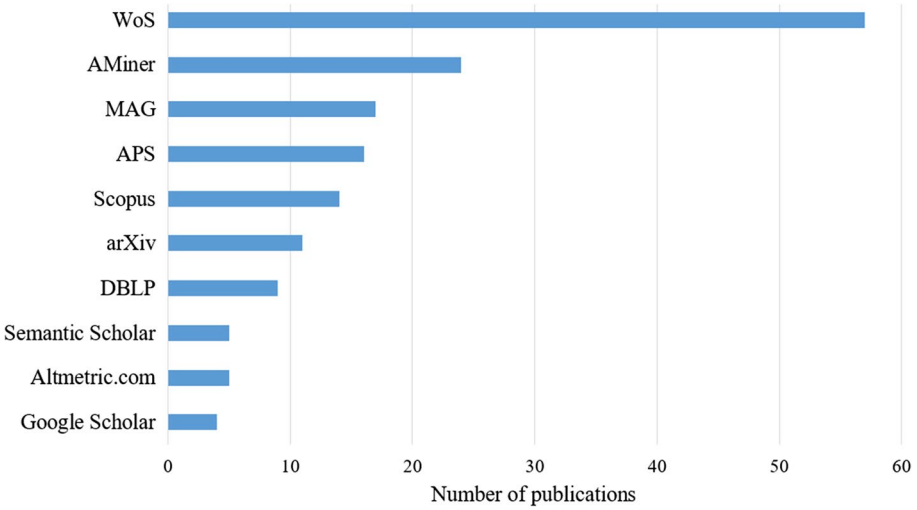

Fig. 10 The top 10 commonly used datasets for scientific impact prediction

arXiv (hep-th)[3] This dataset was released in the KDD cup 2003. It contains over 27,000 articles with 350,000 references on high energy physics from arXiv.

Semantic Scholar (L.weihs)[4] Semantic Scholar was established by the Allen Institute of Artificial Intelligence in 2015. From the initial collection of 3 million articles in the field of computer science, it has now included more than 200 million documents, covering 19 fields such as economics and management. Weihs and Etzioni (2017) summarized the dataset from 1975 to 2015 for detailed analysis and extracted more features, e.g., *h*-index of various scholars.

Evaluation metrics

Due to the lack of a unified gold standard, it is challenging to evaluate the prediction results. We summarize the common evaluation metrics in Table 5.

Model performance

Table 6 summarizes the performance of common prediction tasks and shows the best prediction results. We conduct model performance analysis and select those models with specific dataset sizes and clear prediction results. But it should be noted that due to different datasets used, the prediction results are not comparable even for the same prediction task.

[3] 3 https://kdd.org/kdd-cup/view/kdd-cup-2003/Data.
[4] 4 https://github.com/Lucaweihs/impact-prediction/.

Open challenges and future research directions

Multisource data fusion

The continued growth of scientific corpora and the increasing importance of nontraditional literature have increasingly enriched data sources. How to integrate additional data sources, such as preprints and commercialization data, is the current challenge (Weis & Jacobson, 2021). In addition, the currently used dataset has the problem of missing citation information (Zhang & Wu, 2021), and a paper may have different citations in various literature databases. It may be beneficial to use some external resources, such as Google Scholar, to collect citation data (Du et al., 2021). In addition, a series of multidimensional complex features can be used for scientific impact prediction. However, most current studies choose features that can be obtained in a relatively simple and fast way, which may result in the omission of some features (Yu et al., 2014). Therefore, more indicators should be considered to improve the performance of the prediction model (Klemiński et al., 2021; Kong et al., 2020; Liu et al., 2020; Ma et al., 2021). For example, altmetrics have been explored for scientific impact prediction (Thelwall & Nevill, 2018), but the application of altmetrics is limited by data coverage (Drongstrup et al., 2020). It may be better to combine altmetrics with other features, e.g., citation data, and use several sources for altmetrics to prevent missing data (Du et al., 2021). Therefore, we need to adopt multisource data fusion methods (Zheng, 2015) to process big scholarly data to solve the problem of missing information and provide a reliable basis for scientific impact prediction. Deep learning can be used to learn the hierarchical features of data through unsupervised training (Zhang et al., 2018b) or combine broad learning (Zhang & Yu, 2018) to integrate different data sources and mine more valuable information. The fusion of multisource data can help prevent researchers from focusing on optimizing citations and promote more accurate and objective predictions.

Interpretability and stability

Machine learning methods have shown their effectiveness in this field and can obtain high accuracy. However, the prediction of scientific impact is not only used for "prediction" purposes but is also used to explore the laws of development behind science and deeply understand and promote scientific research. Machine learning algorithms such as BP neural networks have a 'black box' nature and cannot be used to interpret the relationship between selected features and the number of citations (Ruan et al., 2020). Although statistical learning models can explain the relationship between input variables and scientific impact through correlation, the variables are considered to be independent; however, this assumption is not always true (Thelwall & Nevill, 2018). Moreover, the correlation itself is also unstable, and a large number of false associations caused by sample selection bias will be generated, which leads to unexplainable and unstable models (Zhang et al., 2017). In addition, the datasets used by researchers are often limited to a certain field (Hu et al., 2020; Levitt & Thelwall, 2011; Yu et al., 2014), and the stability of models is unclear when applied to different disciplines or interdisciplinary fields (Bornmann & Daniel, 2010; Zhou

Table 5 The common evaluation metrics for scientific impact prediction

Metric	Formula	Description
ACC	$ACC = \frac{TP+TN}{TP+FP+TN+FN}$	TP: True Positive, FP: False Positive, FN: False Positive, TN: True Negative
F1	$F1 = 2 \cdot \frac{precison \cdot recall}{precision+recall}$	$precision = \frac{TP}{TP+FP}$ $recall = \frac{TP}{TP+FN}$
AUC	–	Area under the ROC curve
R^2	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$	–
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$	–
Root Mean Squared Logarithmic Error (RMSLE)	$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(\hat{y}_i+1) - \log(y_i+1))^2}$	–
Mean Square Error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$	–
Mean Absolute Error (MAE)	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert$	–
Mean Absolute Log-scaled Error (MALSE)	$MALSE = \frac{1}{n}\sum_{i=1}^{n}\lvert\log(\hat{y}_i+1) - \log(y_i+1)\rvert$	–
Mean Squared Logarithmic Error (MSLE)	$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(\hat{y}_i+1) - \log(y_i+1))^2$	–
Mean Absolute Percentage Error (MAPE)	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\lvert\frac{\hat{y}_i-y_i}{y_i}\rvert$	–
Pearson correlation coefficient	$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$	–

Table 5 (continued)

Metric	Formula	Description
Spearman's rank correlation coefficient (SPCC)	$\rho = \frac{\sum_{i=1}^{N}(R_1(p_i)-\overline{R}_1)(R_2(p_i)-\overline{R}_2)}{\sqrt{\sum_{i=1}^{N}(R_1(p_i)-\overline{R}_1)^2(R_2(p_i)-\overline{R}_2)^2}}$	$R_1(p_i)$ and $R_2(p_i)$ are the rankings of p_i in two ranking list. $\overline{R}_1$ and $\overline{R}_2$ are the average rankings of all the papers in two ranking list
Normalized Discounted Cumulative Gain (NDCG)	$DCG@K = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}$	rel_i represents the relevance score of the ith ranked entity. $IDCG$ is an ideal ranking of DCG
	$NDCG@K = \frac{DCG@K}{IDCG@K}$	

Table 6 The performance of common prediction tasks

Object	Task	Input	Output	Dataset	Model	Best performance	References
Paper	Citation count prediction	16 features including journal ranking, number of coauthors, network features and semantic features of the abstract	Number of citations 6 years after publication	Scopus: 223,558 papers	XGBoost	ACC: 0.79 AUC: 0.7	Fronzetti Colladon et al. (2020)
		30 features including document type, paper length, JIF, h-index, early citations, etc	Number of citations 5 years after publication	CSSCI: 49,834 papers	SVR, RF, LR, XGBoost, RNN, KNN, BP neural network	MSE: 2.584 R^2: 0.837 (BP neural network)	Ruan et al. (2020)
		10 features including topic attribute, author attribute, network attribute, journal attribute	Citation counts after 1, 5, 10 years	AMiner: 20,000 papers	LR, KNN, SVR, CART	R^2: 0.786 (CART)	Yan et al. (2011)
		2 features including number of citations in previous years and JIF	Number of citations 9 years after publication	WoS:123,128 papers	Statistical learning	R^2 values above 0.8	Abramo et al. (2019a)
		feature numbers range from 55 to 310 for different years including author number, keyword, reference, h-index and network characteristics	Number of citations 10 years after publication	AMiner: 2,365 papers	CNN	R^2: 0.9236 MSE: 0.0231	Xu et al. (2019)

Table 6 (continued)

Object	Task	Input	Output	Dataset	Model	Best performance	References
	Citation sequence prediction	citation count of the paper from the 0th to kth year ($0 < k < 7$)	Citation count from ($k+1$)th to 14th year	WoS: 175,432 papers	RNN	R^2: 0.8 RMSE < 15	Abrishami and Aliakbary (2019)
	Highly-cited paper prediction	23 features including author number, h-index, fist-cited age, JIF and social media data, etc	Three classification: highly-cited, medium-cited or low-cited	WoS: 617 papers	NB, KNN and RF	Precision: 0.947 (RF)	Wang et al. (2019c)
		12 features including journal, author characteristics and keyword popularity, etc	Binary classification: highly-cited or less highly-cited	WoS: 746 papers	ANN, C4.5, LR, SVM	AUC: 0.870 (LR)	Hu et al. (2020)
Scholar	H-index or citation count	70 features including author, paper, journal, network characteristics	Citation counts and h-index after 10 years	arXiv: 39,371 authors	Neural network	Correlation coefficient: 0.728 ± 0.006 (for citation counts prediction) R^2: 0.857 ± 0.004 (for h-index prediction)	Mistele et al. (2019)
		35 features including author, paper, journal, institution characteristics	H-index for a given year	MAG: 79,321 scholars APS: 80,360 scholars	XGBoost, LR, GBDT, CART	ACC: 0.86, R^2: 0.99, MSE: 1.19, MAPE: 0.07, MAE: 0.79 (XGBoost)	Kong et al. (2020)
	Rising star	17 features including publication, scholar, social characteristics	Whether an author is rising star or not	AMiner: 160,423 scholars	KNN, RF, SVM, GBDT, XGBoost	F1 scores close to 0.8	Nie et al. (2019)

Table 6 (continued)

Object	Task	Input	Output	Dataset	Model	Best performance	References
		11 features including publication, scholar, social characteristics	Whether an author is rising star or not	DBLP and AMiner: 44,167 scholars	Maximum entropy Markov model (MEMM), CART, Bayes network, NB	F1 score at least 0.9 (MEMM)	Daud et al. (2015)
		Citation network, coauthor network, paper-venue network, paper author network	Future ranking	MAG: 79,321 scholars	PageRank-like algorithm	Pearson correlation coefficient:0.736	Zhang et al. (2018a)
Journal	JIF	4 features including JIF slope and intercept, percentage of review articles, journal guide	JIF in a given year	WoS: 30 journals	Statistical learning	R^2: 0.993	Valderrama et al. (2018)
Institution	Paper acceptance at the institutional level	19 features including GDP, Q value, AIF value, distance, etc	Number of the accepted papers in the next year from each institution	MAG: 33,953 authors with 19,343 papers from 4,524 institutions	GBDT and XGBoost	NDCG @20: 0.945 (XGBoost)	Bai et al. (2017b)
		7 features including affiliation score, network centrality, author collaboration score, author publishing probability, etc	Number of the accepted papers in the next year from each institution	MAG: 321 institutions	RF	NDCG@20: 0.865,	Wang et al. (2021b)

Table 6 (continued)

Object	Task	Input	Output	Dataset	Model	Best performance	References
Multiple academic entities	Future citations of papers and authors	heterogeneous network of papers, authors and venues	Citation counts of papers and authors after 20 years	APS: 11,475 papers and 14,318 authors	GNN	MSLE: 0.268, ACC: 0.6977 (for papers) MSLE: 0.590 ACC: 0.5162 (for authors)	Zhou et al. (2020a)
	Future rank of papers, researchers and venues	Heterogeneous network of papers, authors and venues	Paper citation counts, author weighted citation count and venue weighted citation count for 1, 2,..., 10 years ahead	ACL:19,891 papers, 15,379 authors and 372 venues	PageRank-like algorithm	Pearson correlation coefficient: 0.817 (for papers), 0.898 (for authors), 0.830 (for venues)	Zhang and Wu (2020)

et al., 2020b). In the future, causal inference (Cui et al., 2020) can be added to machine learning to remove the influence of confounding factors and select meaningful features to improve the generalization ability of prediction models and better explain the driving factors behind academic influence. To test the generality and stability, further studies in various disciplines are needed (Kong et al., 2020; Wen et al., 2020).

Citation content analysis

Due to the difference in the citation motivation, the citation behavior is completely random. Some citations are deep, some are superficial, some are positive and some are negative, and there are even coercive citations and padded citations (Fong & Wilhite, 2017). Most current citation-based prediction studies treat all citations equally; thus, neither the true intention of the citing author nor the different influences of the citing paper can be presented (Giuffrida et al., 2019; Zhang et al., 2019). One possible solution is to divide citations into different levels according to the relevance between the cited and the citing paper (F. Zhang et al., 2019). Another approach is to design relevant metrics to measure citation differences. For example, citation strength, which is usually measured by the number of times a paper is cited in the same article, can be used for evaluating paper influence and author influence (Wan & Liu, 2014), and further studies using machine learning methods can be completed for content-based citation strength estimation (Zhang & Wu, 2021). Moreover, content-based citation analysis has received widespread attention (Ding et al., 2014). With the development of NLP, researchers can obtain the sentiment value of the citation through semantic analysis of the citation text and achieve a deeper and more accurate understanding of the content of the paper (Porwal & Devare, 2020). In future research, transformer models such as BERT and RoBERTa can be applied to enhance the prediction effectiveness (Ma et al., 2021; Wang et al., 2021a).

Dynamic evolution of academic networks

The dynamic changes of scientific impact can be reflected in the academic network. In current studies, although the time evolution is taken into account, for example, new papers are given more weight than old papers, the dynamics of academic influence cannot be sufficiently revealed. It would be better to model the process of network evolution (Kanellos et al., 2021). The link evolution and topological structure change in temporal complex networks provide insights into future works (Bütün & Kaya, 2019). In addition, the network-based approach must face the "cold start" problem (Zhou et al., 2021), which refers to those newly published papers or new scholars, of which there are few links in the academic network. To solve the problem of predicting the influence of new papers or scholars, it is possible to combine network-based methods and metadata information such as topics, authors, and institutions (Zhou et al., 2021). However, how to integrate additional information into academic networks and ranking frameworks is still a challenge (Zhang & Wu, 2021). The application of graph-based machine learning methods such as GNN can simultaneously consider the features of each node in the graph and the features of its neighbors, which can be used to solve the cold start problem and realize the impact prediction of new papers or new authors (Weis & Jacobson, 2021). Future research can pay more attention to embedding methods, specifically for academic networks (Klemiński et al., 2021), and consider the dynamic GNN method (Skarding et al., 2021), which can represent the structure

and timing information of the academic network and capture the dynamic evolution characteristics of academic influence.

Benchmarks and evaluations

Constructing a unified baseline is conducive to the continuation of scientific research in previous findings. Machine learning methods require standards to assess the effectiveness of models. Most of the current studies only compare different ML algorithms to obtain the best model and lack comparison with other papers (Akella et al., 2021; Brizan et al., 2016). Very few papers have made comparison with other scholars' models, which are based on different datasets; the advantages of the proposed method cannot be reflected well (Mistele et al., 2019). Without a standardized dataset, it is difficult to form a unified evaluation standard, and it will be difficult to compare different methods. Many forecasting methods claim to have high accuracy, but they are not effective when applied to different datasets (García-Pérez, 2013). Building benchmarks for scientific impact prediction is an important problem that urgently needs to be solved (Bai et al., 2020). In the future, more evaluation experiments, which should be reproducible on different datasets of different sizes, are needed (Liu et al., 2020). We should actively promote data sharing, strengthen collaborations between researchers and publishing institutions, formulate the open sharing standardization of data and policies, increase researchers' enthusiasm for data sharing, and work together to build benchmarks to promote the rapid development of this research field.

Conclusion

For years, researchers have been attempting to find models that can accurately predict the future impact of academic papers, scholars, publication venues and institutions. In this paper, we conducted a comprehensive review of the literature on predicting future scientific impacts, focusing on prediction tasks, features and methods. There are always many controversies and prejudices on how to measure academic influence. It has been argued that these predictive models will only serve to perpetuate existing academic biases (Chawla, 2021). Therefore, scientific impact prediction should first focus on providing more choices for scientific research and helping researchers discover directions with greater influence in the future in advance. Second, it can be used to identify the driving forces of science and develop predictive models to capture the evolution of technological innovation and better accelerate it. Third, analyzing the mechanism for successfully predicting scientific impact can help design policies that improve scientific enterprise (Fortunato et al., 2018), rather than only focusing on predicting results.

Funding This research is partially supported by the National Natural Science Foundation of China (Grant Nos. 62202395 and 62176221), Sichuan Academic Achievement Analysis and Application Research Center (Grant No. XSCG2021-021), Natural Science Foundation of Sichuan Province (Grant No. 2022NSFSC0930), Fundamental Research Funds for the Central Universities (Grant No. 2682022CX067) and Youth Talent Startup Grant of SWJTU-Leeds Joint School.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics, 5*(4), 594–607.

Abramo, G., D'Angelo, C. A., & Felici, G. (2019a). Predicting publication long-term impact through a combination of early citations and journal impact factor. *Journal of Informetrics, 13*(1), 32–49.

Abramo, G., D'Angelo, C. A., & Reale, E. (2019b). Peer review versus bibliometrics: Which method better predicts the scholarly impact of publications? *Scientometrics, 121*(1), 537–554.

Abrishami, A., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics, 13*(2), 485–499.

Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Predicting scientific success. *Nature, 489*(7415), 201–202.

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), 101128.

Ashton, S. V., & Oppenheim, C. (1978). A method of predicting Nobel prizewinners in chemistry. *Social Studies of Science, 8*(3), 341–348.

Ayaz, S., Masood, N., & Islam, M. A. (2018). Predicting scientific impact based on h-index. *Scientometrics, 114*(3), 993–1010.

Bai, X., Liu, H., Zhang, F., Ning, Z., Kong, X., Lee, I., & Xia, F. (2017a). An overview on evaluating and predicting scholarly article impact. *Information, 8*(3), 73.

Bai, X., Zhang, F., Hou, J., Xia, F., Tolba, A., & Elashkar, E. (2017b). Implicit multi-feature learning for dynamic time series prediction of the impact of institutions. *IEEE Access, 5*, 16372–16382.

Bai, X., Pan, H., Hou, J., Guo, T., Lee, I., & Xia, F. (2020). Quantifying success in science: An overview. *IEEE Access, 8*, 123200–123214.

Bai, X., Zhang, F., & Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics, 13*(1), 407–418.

Bento, C., Martins, B., & Calado, P. (2013). Predicting the Future Impact of Academic Publications. In L. Correia, L. P. Reis, & J. Cascalho (Eds.), *Portuguese Conference on Artificial Intelligence* (pp. 366–377). Springer.

Bertsimas, D., Brynjolfsson, E., Reichman, S., & Silberhoz, J. (2013). Network analysis for predicting academic impact. *Proceedings of the 34th International Conference on Information Systems (ICIS),* 92.

Bhat, H. S., Huang, L.-H., Rodriguez, S., Dale, R., & Heit, E. (2015). Citation prediction using diverse features. *IEEE International Conference on Data Mining Workshop (ICDMW), 2015*, 589–596.

Bornmann, L., & Daniel, H.-D. (2010). Citation speed as a measure to predict the attention an article receives: An investigation of the validity of editorial decisions at Angewandte Chemie International Edition. *Journal of Informetrics, 4*(1), 83–88.

Bornmann, L., Leydesdorff, L., & Wang, J. (2014). How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *Journal of Informetrics, 8*(1), 175–180.

Brizan, D. G., Gallagher, K., Jahangir, A., & Brown, T. (2016). Predicting citation patterns: Defining and determining influence. *Scientometrics, 108*(1), 183–200.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology, 57*(8), 1060–1072.

Bütün, E., & Kaya, M. (2019). Predicting citation count of scientists as a link prediction problem. *IEEE Transactions on Cybernetics, 50*(10), 4518–4529.

Cao, X., Chen, Y., & Liu, K. R. (2016). A data analytic approach to quantifying scientific impact. *Journal of Informetrics, 10*(2), 471–484.

Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2014). Towards a stratified learning approach to predict future citation counts. *IEEE/ACM Joint Conference on Digital Libraries*, 351–360.

Chawla, D. S. (2021). Frosty reception for algorithm that predicts research papers' impact. *Nature*.

Cheang, B., Chu, S. K. W., Li, C., & Lim, A. (2014a). A multidimensional approach to evaluating management journals: Refining PageRank via the differentiation of citation types and identifying the roles that management journals play. *Journal of the Association for Information Science and Technology, 65*(12), 2581–2591.

Cheang, B., Chu, S. K. W., Li, C., & Lim, A. (2014b). OR/MS journals evaluation based on a refined PageRank method: An updated and more comprehensive review. *Scientometrics, 100*(2), 339–361.

Cheang, B., Li, C., Lim, A., & Zhang, Z. (2015). Identifying patterns and structural influences in the scientific communication of business knowledge. *Scientometrics, 103*(1), 159–189.

Chen, C. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology, 63*(3), 431–449.

Cui, P., Shen, Z., Li, S., Yao, L., Li, Y., Chu, Z., & Gao, J. (2020). Causal inference meets machine learning. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3527–3528.

Cummings, D., & Nassar, M. (2020). Structured citation trend prediction using graph neural networks. *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3897–3901.

Danell, R. (2011). Can the quality of scientific work be predicted using information on the author's track record? *Journal of the American Society for Information Science and Technology, 62*(1), 50–60.

Daud, A., Aljohani, N. R., Abbasi, R. A., Rafique, Z., Amjad, T., Dawood, H., & Alyoubi, K. H. (2017). Finding rising stars in co-author networks via weighted mutual influence. *Proceedings of the 26th International Conference on World Wide Web Companion*, 33–41.

Daud, A., Abbasi, R., & Muhammad, F. (2013). Finding rising stars in social networks. In W. Meng, L. Feng, S. Bressan, W. Winiwarter, & W. Song (Eds.), *International conference on database systems for advanced applications* (pp. 13–24). Springer.

Daud, A., Ahmad, M., Malik, M. S. I., & Che, D. (2015). Using machine learning techniques for rising star prediction in co-author network. *Scientometrics, 102*(2), 1687–1711.

Davletov, F., Aydin, A. S., & Cakmak, A. (2014). High impact academic paper prediction using temporal and topological features. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 491–498.

de Abreu Batista-Jr, A., Gouveia, F. C., & Mena-Chalco, J. P. (2021). Predicting the Q of junior researchers using data from the first years of publication. *Journal of Informetrics, 15*(2), 101130.

Dey, R., Roy, A., Chakraborty, T., & Ghosh, S. (2017). Sleeping beauties in computer science: Characterization and early identification. *Scientometrics, 113*(3), 1645–1663.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology, 65*(9), 1820–1833.

Dong, Y., Johnson, R. A., & Chawla, N. V. (2016). Can scientific impact be predicted? *IEEE Transactions on Big Data, 2*(1), 18–30.

Drongstrup, D., Malik, S., Aljohani, N. R., Alelyani, S., Safder, I., & Hassan, S.-U. (2020). Can social media usage of scientific literature predict journal indices of AJG, SNIP and JCR? An altmetric study of Economics. *Scientometrics, 125*(2), 1541–1558.

Du, W., Xie, Z., & Lv, Y. (2021). Predicting publication productivity for authors: Shallow or deep architecture? *Scientometrics, 126*(7), 5855–5879.

Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLoS ONE, 12*(12), e0187394.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., & Uzzi, B. (2018). Science of science. *Science, 359*(6379), eaao0185.

Fronzetti Colladon, A., D'Angelo, C. A., & Gloor, P. A. (2020). Predicting the future success of scientific publications through social network and semantic analysis. *Scientometrics, 124*(1), 357–377.

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.

García-Pérez, M. A. (2013). Limited validity of equations to predict the future h index. *Scientometrics, 96*(3), 901–909.

Gardfield, E. (1977). The 250 "most-cited primary authors, 1961–1975" Part II: The correlation between citedness, noble prizes and academy memberships. *Current Comments, 50*, 5–15.

Gingras, Y., & Wallace, M. (2010). Why it has become more difficult to predict Nobel Prize winners: A bibliometric analysis of nominees and winners of the chemistry and physics prizes (1901–2007). *Scientometrics, 82*(2), 401–412.

Giuffrida, C., Abramo, G., & D'Angelo, C. A. (2019). Are all citations worth the same? Valuing citations by the value of the citing items. *Journal of Informetrics, 13*(2), 500–514.

Ha, L., Jiang, W., Bi, C., Zhang, R., Zhang, T., & Wen, X. (2016). How online usage of subscription-based journalism and mass communication research journal articles predicts citations. *Learned Publishing, 29*(3), 183–192.

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., & Wilson, S. (2008). What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics, 76*(1), 169–185.

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences, 104*(49), 19193–19198.

Holm, A. N., Plank, B., Wright, D., & Augenstein, I. (2020). Longitudinal citation prediction using temporal graph neural networks. ArXiv Preprint ArXiv: 2012.05742.

Hou, J., Pan, H., Guo, T., Lee, I., Kong, X., & Xia, F. (2019). Prediction methods and applications in the science of science: A survey. *Computer Science Review, 34*, 100197.

Hu, Y.-H., Tai, C.-T., Liu, K. E., & Cai, C.-F. (2020). Identification of highly-cited papers using topic-model-based and bibliometric features: The consideration of keyword popularity. *Journal of Informetrics, 14*(1), 101004.

Ibáñez, A., Larrañaga, P., & Bielza, C. (2011). Predicting the h-index with cost-sensitive naive Bayes. *2011 11th International Conference on Intelligent Systems Design and Applications*, 599–604.

Jensen, P., Rouquier, J.-B., & Croissant, Y. (2009). Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics, 78*(3), 467–479.

Jiang, S., Koch, B., & Sun, Y. (2021). HINTS: Citation time series prediction for new publications via dynamic heterogeneous information network embedding. *Proceedings of the Web Conference, 2021*, 3158–3167.

Kanellos, I., Vergoulis, T., Sacharidis, D., Dalamagas, T., & Vassiliou, Y. (2021). Ranking papers by their short-term scientific impact. *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 1997–2002.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences, 112*(24), 7426–7431.

Klemiński, R., Kazienko, P., & Kajdanowicz, T. (2021). Where should I publish? Heterogeneous, networks-based prediction of paper's citation success. *Journal of Informetrics, 15*(3), 101200.

Klimek, P., Jovanovic, S., Egloff, A., & Schneider, R. (2016). Successful fish go with the flow: Citation impact prediction based on centrality measures for term–document networks. *Scientometrics, 107*(3), 1265–1282.

Kong, X., Zhang, J., Zhang, D., Bu, Y., Ding, Y., & Xia, F. (2020). The gene of scientific success. *ACM Transactions on Knowledge Discovery from Data (TKDD), 14*(4), 1–19.

Laurance, W. F., Useche, D. C., Laurance, S. G., & Bradshaw, C. J. (2013). Predicting publication success for biologists. *BioScience, 63*(10), 817–823.

Lee, D. H. (2019). Predicting the research performance of early career scientists. *Scientometrics, 121*(3), 1481–1504.

Levitt, J. M., & Thelwall, M. (2011). A combined bibliometric indicator to predict article impact. *Information Processing & Management, 47*(2), 300–308.

Li, X.-L., Foo, C. S., Tew, K. L., & Ng, S.-K. (2009). Searching for rising stars in bibliography networks. *International Conference on Database Systems for Advanced Applications*, 288–292.

Li, C.-T., Lin, Y.-J., Yan, R., & Yeh, M.-Y. (2015). Trend-based citation count prediction for research articles. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 659–671.

Li, L., & Tong, H. (2015). The child is father of the man: Foresee the success at the early stage. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 655–664.

Li, S., Zhao, W. X., Yin, E. J., & Wen, J.-R. (2019a). A neural citation count prediction model based on peer review text. *Proceedings of the 2019a Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4914–4924.

Li, W., Aste, T., Caccioli, F., & Livan, G. (2019b). Early coauthorship with top scientists predicts success in academic careers. *Nature Communications, 10*(1), 1–9.

Lindahl, J. (2018). Predicting research excellence at the individual level: The importance of publication rate, top journal publications, and top 10% publications in the case of early career mathematicians. *Journal of Informetrics, 12*(2), 518–533.

Lindahl, J., Colliander, C., & Danell, R. (2020). Early career performance and its correlation with gender and publication output during doctoral education. *Scientometrics, 122*(1), 309–330.

Liu, L., Yu, D., Wang, D., & Fukumoto, F. (2020). Citation count prediction based on neural hawkes model. *IEICE Transactions on Information and Systems, 103*(11), 2379–2388.

Livne, A., Adar, E., Teevan, J., & Dumais, S. (2013). Predicting citation counts using text and graph mining. *Proc. the IConference 2013 Workshop on Computational Scientometrics: Theory and Applications*, 1–4.

Luo, Z., He, J., Qian, J., Wang, Y., Chen, J., & Lu, W. (2020). Can scientific publication's network structural features predict its citation? *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in, 2020*, 485–486.

Ma, A., Liu, Y., Xu, X., & Dong, T. (2021). A deep-learning based citation count prediction model with paper metadata semantic features. *Scientometrics, 126*(8), 6803–6823.

Ma, Y., & Uzzi, B. (2018). Scientific prize network predicts who pushes the boundaries of science. *Proceedings of the National Academy of Sciences, 115*(50), 12608–12615.

Mahalakshmi, G. S., Sendhilkumar, S., Jancy, P., & Easwarakumar, K. S. (2020). A Neural Learning Approach for Prediction of Research Citations Using Article Semantics. *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 816–819.
Mazloumian, A. (2012). Predicting scholars' scientific impact. *PLoS ONE, 7*(11), e49246.
Mistele, T., Price, T., & Hossenfelder, S. (2019). Predicting authors' citation counts and h-indices with a neural network. *Scientometrics, 120*(1), 87–104.
Nezhadbiglari, M., Gonçalves, M. A., & Almeida, J. M. (2016). Early prediction of scholar popularity. *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 181–190.
Nie, Y., Zhu, Y., Lin, Q., Zhang, S., Shi, P., & Niu, Z. (2019). Academic rising star prediction via scholar's evaluation model and machine learning techniques. *Scientometrics, 120*(2), 461–476.
Onodera, N., & Yoshikane, F. (2015). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology, 66*(4), 739–764.
Panagopoulos, G., Tsatsaronis, G., & Varlamis, I. (2017). Detecting rising stars in dynamic collaborative networks. *Journal of Informetrics, 11*(1), 198–222.
Park, H.-M., Sinshaw, Y. B., & Sohn, K.-A. (2017). Temporal citation network-based feature extraction for cited count prediction. *International Conference on Mobile and Wireless Technology*, 380–388.
Penner, O., Pan, R. K., Petersen, A. M., Kaski, K., & Fortunato, S. (2013). On the predictability of future impact in science. *Scientific Reports, 3*(1), 1–8.
Pobiedina, N., & Ichise, R. (2016). Citation count prediction as a link prediction problem. *Applied Intelligence, 44*(2), 252–268.
Põder, E. (2017). A framework for the measurement and prediction of an individual scientist's performance. *Trames, 21*(1), 3–14.
Porwal, P., & Devare, M. H. (2020). Citation Classification Prediction Implying Text Features Using Natural Language Processing and Supervised Machine Learning Algorithms. *International Conference on Recent Trends in Image Processing and Pattern Recognition*, 540–552.
Qian, Y., Dong, Y., Ma, Y., Jin, H., & Li, J. (2016). Feature engineering and ensemble modeling for paper acceptance rank prediction. ArXiv Preprint ArXiv: 1611.04369.
Rokach, L., Kalech, M., Blank, I., & Stern, R. (2011). Who is going to win the next association for the advancement of artificial intelligence fellowship award? Evaluating researchers by mining bibliographic data. *Journal of the American Society for Information Science and Technology, 62*(12), 2456–2470.
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), 101039.
Sandulescu, V., & Chiru, M. (2016). Predicting the future relevance of research institutions-The winning solution of the KDD Cup 2016. ArXiv Preprint ArXiv: 1609.02728.
Sayyadi, H., & Getoor, L. (2009). Futurerank: Ranking scientific articles by predicting their future pagerank. *Proceedings of the 2009 SIAM International Conference on Data Mining*, 533–544.
Schreiber, M. (2013). How relevant is the predictive power of the h-index? A case study of the time-dependent Hirsch index. *Journal of Informetrics, 7*(2), 325–329.
Shen, H., Wang, D., Song, C., & Barabási, A.-L. (2014). Modeling and predicting popularity dynamics via reinforced poisson processes. *Proceedings of the AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v28i1.8739
Shuang, Q. (2016). *Heterogenous Graph Mining for Measuring the Impact of Research Institutions*.
Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science, 354*(6312), aaf5239.
Singh, M., Patidar, V., Kumar, S., Chakraborty, T., Mukherjee, A., & Goyal, P. (2015). The role of citation context in predicting long-term citation profiles: An experimental study based on a massive bibliographic text dataset. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1271–1280.
Skarding, J., Gabrys, B., & Musial, K. (2021). Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. *IEEE Access, 9*, 79143–79168.
Sohrabi, B., & Iraj, H. (2017). The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts. *Scientometrics, 110*(1), 243–251.
Stegehuis, C., Litvak, N., & Waltman, L. (2015). Predicting the long-term citation impact of recent publications. *Journal of Informetrics, 9*(3), 642–657.
Stern, D. I. (2014). High-ranked social science journal articles can be identified from early citation information. *PLoS ONE, 9*(11), e112520. https://doi.org/10.1371/journal.pone.0112520
Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics, 107*(3), 1195–1225.
Thelwall, M., & Nevill, T. (2018). Could scientists use Altmetric: Com scores to predict longer term citation counts? *Journal of Informetrics, 12*(1), 237–248.

Timilsina, M., Davis, B., Taylor, M., & Hayes, C. (2016). Towards predicting academic impact from mainstream news and weblogs: A heterogeneous graph based approach. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016*, 1388–1389.

Valderrama, P., Escabias, M., Jiménez-Contreras, E., Valderrama, M. J., & Baca, P. (2018). A mixed longitudinal and cross-sectional model to forecast the journal impact factor in the field of Dentistry. *Scientometrics, 116*(2), 1203–1212.

Van Dijk, D., Manor, O., & Carey, L. B. (2014). Publication metrics and success on the academic job market. *Current Biology, 24*(11), R516–R517.

Van Noorden, R. (2017). The science That's. *Nature, 552*, 162–164.

Van Raan, A. F. (2004). Sleeping beauties in science. *Scientometrics, 59*(3), 467–472.

Walker, D., Xie, H., Yan, K.-K., & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment, 2007*(06), P06010.

Wan, X., & Liu, F. (2014). Are all literature citations equally important? Automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology, 65*(9), 1929–1938.

Wang, S., Xie, S., Zhang, X., Li, Z., Yu, P. S., & Shu, X. (2014). Future influence ranking of scientific literature. *Proceedings of the 2014 SIAM International Conference on Data Mining*, 749–757.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science, 342*(6154), 127–132.

Wang, F., Fan, Y., Zeng, A., & Di, Z. (2019a). Can we predict ESI highly cited publications? *Scientometrics, 118*(1), 109–125.

Wang, M., Jiao, S., Chai, K.-H., & Chen, G. (2019b). Building journal's long-term impact: Using indicators detected from the sustained active articles. *Scientometrics, 121*(1), 261–283.

Wang, M., Wang, Z., & Chen, G. (2019c). Which can better predict the future success of articles? Bibliometric indices or alternative metrics. *Scientometrics, 119*(3), 1575–1595.

Wang, J., Zhang, F., Li, Y., & Liu, D. (2020). Attention-based multi-fusion method for citation prediction. In J.-S. Pan, J. Li, P.-W. Tsai, & L. C. Jain (Eds.), *Advances in intelligent information hiding and multimedia signal processing* (pp. 315–322). Springer.

Wang, K., Shi, W., Bai, J., Zhao, X., & Zhang, L. (2021a). Prediction and application of article potential citations based on nonlinear citation-forecasting combined model. *Scientometrics, 126*(8), 6533–6550.

Wang, W., Zhang, J., Zhou, F., Chen, P., & Wang, B. (2021b). Paper acceptance prediction at the institutional level based on the combination of individual and network features. *Scientometrics, 126*(2), 1581–1597.

Wang, M., Yu, G., Xu, J., He, H., Yu, D., & An, S. (2012). Development a case-based classifier for predicting highly cited papers. *Journal of Informetrics, 6*(4), 586–599.

Wang, M., Yu, G., & Yu, D. (2011). Mining typical features for highly cited papers. *Scientometrics, 87*(3), 695–706.

Wang, S., Xie, S., Zhang, X., Li, Z., Yu, P. S., & He, Y. (2016). Coranking the future influence of multiobjects in bibliographic network through mutual reinforcement. *ACM Transactions on Intelligent Systems and Technology (TIST), 7*(4), 1–28.

Way, S. F., Morgan, A. C., Clauset, A., & Larremore, D. B. (2017). The misleading narrative of the canonical faculty productivity trajectory. *Proceedings of the National Academy of Sciences, 114*(44), E9216–E9223.

Way, S. F., Morgan, A. C., Larremore, D. B., & Clauset, A. (2019). Productivity, prominence, and the effects of academic environment. *Proceedings of the National Academy of Sciences, 116*(22), 10729–10733.

Weihs, L., & Etzioni, O. (2017). Learning to predict citation-based impact measures. *ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017*, 1–10.

Weis, J. W., & Jacobson, J. M. (2021). Learning on knowledge graph dynamics provides an early warning of impactful research. *Nature Biotechnology, 39*(10), 1300–1307.

Wen, J., Wu, L., & Chai, J. (2020). Paper citation count prediction based on recurrent neural network with gated recurrent unit. *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 303–306.

Wilson, J., Mohan, R., Arif, M., Chaudhury, S., & Lall, B. (2016). Ranking academic institutions on potential paper acceptance in upcoming conferences. ArXiv Preprint ArXiv: 1610.02828.

Wu, X., Fu, Q., & Rousseau, R. (2008). On indexing in the Web of Science and predicting journal impact factor. *Journal of Zhejiang University Science B, 9*(7), 582–590.

Wu, Z., Lin, W., Liu, P., Chen, J., & Mao, L. (2019). Predicting long-term scientific impact based on multifield feature extraction. *IEEE Access, 7*, 51759–51770.

Xia, F., Wang, W., Bekele, T. M., & Liu, H. (2017). Big scholarly data: A survey. *IEEE Transactions on Big Data, 3*(1), 18–35.

Xiao, C., Han, J., Fan, W., Wang, S., Huang, R., & Zhang, Y. (2019). Predicting scientific impact via heterogeneous academic network embedding. *Pacific Rim International Conference on Artificial Intelligence*, 555–568.

Xiao, C., Sun, L., Han, J., & Qiao, Y. (2021). Heterogeneous academic network embedding based multivariate random-walk model for predicting scientific impact. *Applied Intelligence*, 1–18.
Xie, Z. (2020). Predicting publication productivity for researchers: A piecewise Poisson model. *Journal of Informetrics, 14*(3), 101065.
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, 7*, 92248–92258.
Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011). Citation count prediction: Learning to estimate future citations for literature. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 1247–1252.
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.
Yu, X., Szymanski, B. K., & Jia, T. (2021). Become a better you: Correlation between the change of research direction and the change of scientific performance. *Journal of Informetrics, 15*(3), 101193.
Yuan, S., Tang, J., Zhang, Y., Wang, Y., & Xiao, T. (2018). Modeling and predicting citation count via recurrent neural network with long short-term memory. ArXiv Preprint ArXiv: 1811.02129.
Zhang, C., Liu, C., Yu, L., Zhang, Z.-K., & Zhou, T. (2017). Identifying the academic rising stars via pairwise citation increment ranking. *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*, 475–483.
Zhang, F., Bai, X., & Lee, I. (2019). Author impact: Evaluations, predictions, and challenges. *IEEE Access, 7*, 38657–38669.
Zhang, F., & Wu, S. (2020). Predicting future influence of papers, researchers, and venues in a dynamic academic network. *Journal of Informetrics, 14*(2), 101035.
Zhang, F., & Wu, S. (2021). Measuring academic entities' impact by content-based citation analysis in a heterogeneous academic network. *Scientometrics, 126*(8), 7197–7222.
Zhang, J., Ning, Z., Bai, X., Wang, W., Yu, S., & Xia, F. (2016a). Who are the rising stars in academia? *IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2016*, 211–212.
Zhang, J., Xia, F., Wang, W., Bai, X., Yu, S., Bekele, T. M., & Peng, Z. (2016b). Cocarank: A collaboration caliber-based method for finding academic rising stars. *Proceedings of the 25th International Conference Companion on World Wide Web*, 395–400.
Zhang, J., Xu, B., Liu, J., Tolba, A., Al-Makhadmeh, Z., & Xia, F. (2018a). PePSI: Personalized prediction of scholars' impact in heterogeneous temporal academic networks. *IEEE Access, 6*, 55661–55672.
Zhang, L., Xie, Y., Xidao, L., & Zhang, X. (2018b). Multi-source heterogeneous data fusion. *International Conference on Artificial Intelligence and Big Data (ICAIBD), 2018*, 47–51.
Zhang, X., Fu, L., & Wang, X. (2018c). Ranking the Future Influence of Scientific Literatures. *2018c IEEE 4th International Conference on Computer and Communications (ICCC)*, 2362–2371.
Zhang, J., & Yu, P. S. (2018). Broad learning: An emerging area in social network analysis. *ACM SIGKDD Explorations Newsletter, 20*(1), 24–50.
Zheng, Y. (2015). Methodologies for cross-domain data fusion: An overview. *IEEE Transactions on Big Data, 1*(1), 16–34.
Zhou, F., Xu, X., Li, C., Trajcevski, G., Zhong, T., & Zhang, K. (2020a). A heterogeneous dynamical graph neural networks approach to quantify scientific impact. ArXiv Preprint ArXiv: 2003.12042.
Zhou, Y., Cheng, H., Li, Q., & Wang, W. (2020b). Diversity of temporal influence in popularity prediction of scientific publications. *Scientometrics, 123*(1), 383–392.
Zhou, Y., Wang, R., Zeng, A., & Zhang, Y.-C. (2020c). Identifying prize-winning scientists by a competition-aware ranking. *Journal of Informetrics, 14*(3), 101038.
Zhou, W., Gu, J., & Jia, Y. (2018). H-Index-based link prediction methods in citation network. *Scientometrics, 117*(1), 381–390.
Zhou, Y., Li, Q., Yang, X., & Cheng, H. (2021). Predicting the popularity of scientific publications by an age-based diffusion model. *Journal of Informetrics, 15*(4), 101177.
Zoller, D., Doerfel, S., Jäschke, R., Stumme, G., & Hotho, A. (2016). Posted, visited, exported: Altmetrics in the social tagging system BibSonomy. *Journal of Informetrics, 10*(3), 732–749.
Zuo, Z., & Zhao, K. (2021). Understanding and predicting future research impact at different career stages—A social network perspective. *Journal of the Association for Information Science and Technology, 72*(4), 454–472.