Investigating the contribution of author- and publication-specific features to scholars' h-index prediction

Momeni, Fakhri; Mayr, Philipp; Dietze, Stefan

This work has been funded by the Federal Ministry of Education and Research of Germany (BMBF) (grant no.: 01PU17005A)

Momeni et al. *EPJ Data Science* (2023) 12:45
https://doi.org/10.1140/epjds/s13688-023-00421-6

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Investigating the contribution of author- and publication-specific features to scholars' h-index prediction

Fakhri Momeni[1*], Philipp Mayr[1] and Stefan Dietze[1,2]

*Correspondence:
fakhri.momeni@t-online.de
[1]GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany
Full list of author information is available at the end of the article

Abstract

Evaluation of researchers' output is vital for hiring committees and funding bodies, and it is usually measured via their scientific productivity, citations, or a combined metric such as the h-index. Assessing young researchers is more critical because it takes a while to get citations and increment of h-index. Hence, predicting the h-index can help to discover the researchers' scientific impact. In addition, identifying the influential factors to predict the scientific impact is helpful for researchers and their organizations seeking solutions to improve it. This study investigates the effect of the author, paper/venue-specific features on the future h-index. For this purpose, we used a machine learning approach to predict the h-index and feature analysis techniques to advance the understanding of feature impact. Utilizing the bibliometric data in Scopus, we defined and extracted two main groups of features. The first relates to prior scientific impact, and we name it 'prior impact-based features' and includes the number of publications, received citations, and h-index. The second group is 'non-prior impact-based features' and contains the features related to author, co-authorship, paper, and venue characteristics. We explored their importance in predicting researchers' h-index in three career phases. Also, we examined the temporal dimension of predicting performance for different feature categories to find out which features are more reliable for long- and short-term prediction. We referred to the gender of the authors to examine the role of this author's characteristics in the prediction task. Our findings showed that gender has a very slight effect in predicting the h-index. Although the results demonstrate better performance for the models containing prior impact-based features for all researchers' groups in the near future, we found that non-prior impact-based features are more robust predictors for younger scholars in the long term. Also, prior impact-based features lose their power to predict more than other features in the long term.

Keywords: h-index prediction; Feature importance; Academic mobility; Machine learning; Open access publishing

1 Introduction

Predicting scientific impact helps to anticipate the career trajectories of researchers and reveal mechanisms of the scientific process that influence future impact, which has al-

Springer

ways been a concern of individual researchers, universities, recruitment committees, and funding agencies. Also, it can reveal factors influencing the future outcome and propose path-ways to young researchers on how to improve future impact and their organizations for more support.

Scientific productivity and received citations are the basis for many evaluation metrics (e.g., h-index [1], g-index [2], h_s-index [3]). The h-index is the most common metric which evaluates the scholars' scientific impact since it measures researchers' productivity and citation impact and has a leading role in hiring and funding decisions. Therefore, predicting this metric is crucial for these purposes. The shorter publication record, received citations, and h-index (prior impact-based features) simplify the h-index prediction task because these features reflect the scholar's impact. Since more senior scholars have a distinguished research profile, predicting their h-index is easier. Assessing the future impact is more pivotal for young scholars than seniors because prior impact-based features are less available for junior researchers as they have a shorter data history. The prediction task will be more complicated for rising stars (who have a lower research profile at the beginning of their career compared to other authors in the same career stage but may become prominent contributors in the future [4]), and we need non-prior impact-based features to evaluate their impact in the long term. Although previous studies demonstrated high accuracy by employing prior impact-based features [5–7], they displayed a substantial decline in the performance of predicting the h-index in the distant future. We hypothesise that publication/citation-based features may be efficient short-term predictors, but other feature categories may be more efficient in predicting long-term impact.

To address these limitations and improve the accuracy of h-index prediction, this study takes a comprehensive approach by investigating a wide array of features and feature combinations. We consider traditional publication/citation-based features and explore other feature categories that may play a role in predicting long-term impact. Our primary objective is to gain a deeper understanding of feature contributions to the h-index prediction task for researchers at different career stages. Our investigation involves analyzing various features and feature combinations in the context of h-index prediction. Drawing from prior research associating specific features with productivity and received citations, we examine how these attributes contribute to researchers' future h-index. To accomplish this, we leverage a machine learning approach to predict the h-index for the upcoming ten years and conduct an extensive feature analysis. To assess the temporal stability of our predictions, we implement our method on three distinct groups of authors: junior, middle-level, and senior researchers. By comparing the accuracy of different feature combinations within each group, we gain insights into the efficacy of the predictive models over time.

In summary, our study makes three significant contributions to the field:

1. *Feature impact analysis:* We advance the understanding of the impact of different feature categories on various h-index prediction tasks for researchers in different career phases and examine the reliability of these predictions.
2. *Temporal dimension of feature performance:* We investigate the temporal dimension of predictors to advance the understanding of feature performance depending on the time window considered for the future prediction, i.e., to understand which features/categories perform better for long- and short-term prediction regarding their seniority.

3. *Novel features:* We introduce and investigate the effect of non-prior impact-based features, namely gender and academic mobility, on the prediction task to reveal the influential factors on the scientific impact (prior impact-based features that implicitly or explicitly encode citation counts simplify the h-index prediction task dramatically by providing the model with data that directly influences the target metric (h-index)).

2 Related work

To identify the future scientific impact, several studies focus on predicting the citations count for a specific paper [8–12], others tried to predict the impact at the author level with the h-index [5–7, 13]. Among all models and methods presented in these studies to predict the h-index, those that took the number of prior publications, received citations, or the current h-index (prior impact-based features) into consideration achieved the highest performance. Although prior impact-based features are the strongest predictors of future impact, sometimes we need to predict it using the other author, paper, and venue characteristics.

2.1 Features used for the prediction tasks

Many studies employed various properties of papers, venues, authors, and their coauthors to predict the scientific impact. Abrishami and Aliakbary [9] and Bai et al. [8] use time series methods and early citations count to predict the number of citations in the long term. Jiang et al. [10] presented a citation time series approach to predict the citations for newly published papers. They used the paper's topic (via keyword), author reputation, venue prestige, and temporal cues (e.g., increasing network centrality over time) to detect citation signals and convert them into signals for citation time series generation. Nie et al. [14] utilized some features and categorized them into the author (regarding citations and publication), venue, social (coauthor), and temporal (average citation increment of the author and coauthors within two years) features and examined their importance in predicting academic rising star. Ayaz et al. [5] and Weihs and Etzioni [6] used the number of current publications, citations, or h-index with other features to predict the future h-index and both presented models with $R^2 = 0.93$. Wu et al. [7] included related indicators to these features, such as changes in citations and h-index over the last two years to the predictors' list and demonstrated a model with a higher precision $R^2 = 0.97$. Further studies focused on other feature types rather than prior impact-based features to identify the influential factors on the scientific impact of researchers. For example, McCarty et al. [15] investigated the relationship between some characteristics of the coauthor network and the h-index. Their results showed the significance of coauthors' productivity via collaborating with many authors and their impact on predicting the h-index. Nikolentzos et al. [13] extracted two types of features, papers' textual content and graph features (related to collaboration patterns), and found that graph features alone are more robust predictors. Dong et al. [16] studied the contribution of a publication to the author's h-index and found that topical authority and publication venues are the most predictive features in the absence of citation-related features of prior publications. Otherwise, they reported citation count as the most decisive factor in predicting the future h-index. Jiang et al. [10] found that certain features, such as the author's reputation, are more predictive than others. Therefore, they applied trainable weights to preserve the unequal contribution of different kinds of features. Ayaz et al. [5] reported the career age, number of high-quality

papers, and number of publications in distinct journals as the most compelling feature in predicting the h-index after prior impact-based features. They observed a lower performance for younger researchers and concluded the investigated features are insufficient to predict their h-index and a need to evaluate future features for better prediction.

Wu et al. [7] investigated the stability of predictive models for long-term prediction (ten future years) and compared their method with state-of-the-art [5, 6, 16]. They used time series features (the history of the h-index) and more impact-based features in their analyses, which are less valuable to predict the future impact of young researchers. They found better performance among all mentioned works. However, they included only the authors with an h-index higher than four and junior researchers whose predicting their scientific impact is more challenging have been excluded from their study.

We tackle these issues by investigating novel author- and paper-specific features for the prediction task and verifying their contribution to the h-index prediction for researchers with varying scientific experiences.

2.2 Influential factors on scientific impact

In the following, we categorize the features affecting the scientific impacts into three groups: demographic, paper/venue, and coauthor-based factors, and report the previous related studies.

2.2.1 Demographic factors

Academic mobility In contemporary science, collaboration plays a significant role, and international academic mobility affects the collaboration networks, which furthers knowledge transmission among countries and scholars. Therefore, many studies have focused on investigating its impact on science and scientists. Our recent study [17] revealed the positive impact of international mobility on the number of publications and received citations. However, mobile researchers do not necessarily perform better than those without mobility experience. Singh [18] found that differences in research outputs between returnee Ph.D. holders and those trained in their home country are field-specific and depend on their seniority. Netz et al. [19] reviewed the studies that investigated the effect of mobility on some scientific outcomes and found that most studies suggest a positive effect on mobility. But they reported some studies that demonstrated a negative effect on productivity and citation impact and proposed a positive impact of mobility only under specific circumstances. Liu et al. [20] found that international collaboration before mobility has an essential role in high performance after mobility. The reputation of institutions is another influential factor they discovered in their study.

Gender Gender differences in science and scientific impact have been the subject of many studies in various fields. A new study on the Breast Surgery Fellowship Faculty [21] found no noticeable gender difference between assistant professors but a higher h-index for men professors than women. [22] studied the gender gap in social sciences and found the difference in all career phases, especially in full professor positions. In contrast, the study's results by Lopez et al. [23] demonstrated a higher h-index for men among academic ophthalmologists. Still, controlling the range of publications, they found the same or more impact for women in the later career phases. The results of the study by Kelly et al. [24] indicated that although the h-index of men is higher than women for ecologists

and evolutionary biologists, there is no gender difference in the h-index once we control for publication rate. However, other studies [25, 26] examined the relationship between received citations and funding available from Web of Science data and found a weak correlation between them.

Income level In many countries, governments are the primary source of financial support for scientific progress. Gantman [27] demonstrated the positive effect of economic development on scientific productivity in all scientific fields. Confraria et al. [28] displayed a U-shape relationship between Gross Domestic Product (GDP) per capita and received citations and found the citation impact correlates positively with the nation's wealth after a certain GDP per capita level. However, their results showed that international collaboration is crucial for higher citation impact among all countries.

2.2.2 Paper and venue factors

Scientific field The average scholars' h-index of researchers differs among fields because productivity and the rate of citing vary from one to another [29, 30]. Iglesias and Pecharrom [31] showed the varying ranges of the h-index across fields and suggested a multiplicative correction to the h-index based on the scientific field to compare the scientists' research impact from different areas.

Journal quality Reputable journals increase the visibility of papers and the probability of receiving citations. Petersen and Penner [32] found that publishing in high-quality journals decreases the average time interval between the author's future publications in those journals and has a cumulative citation advantage for the author.

Open access Free access to publications in online form increases the probability of reading and citing papers. Various studies investigated the Open Access Citation Advantage (OACA), and most found a positive effect on received citations [33–36]. Langham-Putrow et al. [37] did a systematic review of the OACA and reported that among 143 studies, 47.8% confirmed OACA, 37% found no OACA, and 24% found OACA for a subset of their sample. Also, the result of our recent study [38] showed substantially higher citations for preprint papers, making publications freely available. Momeni et al. [39] examined the association of open access publishing with received citations and found a higher percentage of highly cited papers published in the open-access model than those in the closed-access model.

2.2.3 Coauthor factors

The number of the paper's citations received reveals the scientific impact of all authors, and hence it can vary according to their collaboration pattern. Hsu and Huang [40] found a positive correlation between the number of coauthors and received citations. Also, the result of the study by Puuska et al. [41] showed fewer citation scores for single-authored publications. Sarigöl et al. [42] tried to predict highly cited papers via the centrality of their authors in the co-authorship network and found a positive correlation between highly cited publications and highly centralized authors.

Other studies [41, 43] examined the citation impact of international coauthors and demonstrated a positive relation between international collaboration and received citations.

2.3 Prediction approaches

Many studies employed machine learning regression and classification approaches to predict the scientific impact of publications and researchers [6, 7, 9–11, 13]. The most common methods in these studies were regression models such as Support Vector Regression (SVR), Gradient Boosted Regression Trees (GBRT) or Gradient Boosting (GB), Gradient-Boosting Decision Tree (GBDT), Extreme Gradient Boosting (XGBoost), Random Forest (RF), K-nearest Neighbour (KNN), and Neural Networks (NN). Nie et al. [14] introduced a classification method to detect the academic rising stars (who have a lower research profile at the beginning of their career compared to other authors in the same career stage but may become prominent contributors in the future) and found better performance for KNN algorithm for small datasets, but a relatively stable result for GBDT, GB, RF, and RF with the change of dataset size. Ruan et al. [11] examined the performance of different regression algorithms and reported the best performance for Backpropagation neural network. Wu et al. [7] examined SVR, RF, GBRT, and XGBoost regression models for h-index prediction and obtained the best performance for XGBoost. The performance of methods for predicting the h-index in different ranges depends on applied features. By using prior impact-based features and regression models, previous studies [5–7] presented models with $R^2 > 0.90$ for the first predicting year and decreased in the next predicting years. However, none of these studies investigated the extent of the contribution of different features in the prediction task. Our study examines the contribution of features to the h-index prediction via feature selection/ranking approaches to understanding the influential factors better.

3 Data and methods

3.1 Describing the dataset

We used the in-house Scopus database maintained by the German Competence Centre for Bibliometrics (Scopus-KB), 2020 version, as the central resource of analyses and employed Scopus author Id to identify authors. We defined the career age of authors by the years between the first and last publication time. We took authors who started publishing after 1994 and used their publications until 2008 to calculate the features' value. We detected the gender status of authors by a combined name and image-based approach introduced by Karimi et al. [44], which results in a binary variable. We acknowledge that a person's gender can not be split into male and female, and if we consider the social dimensions, we have more gender identities.

To remove "not active authors" from the analyzed data, we included just those authors who had at least five years of career age, an h-index higher than zero and matched the threshold of one publication per three years in their career age. Excluding authors without gender status results in a final list of 1,824,203 authors. Table 1 presents some information about the distribution of analysed papers among main research domains (categorized by the All Science Journal Classification (ASJC) System in Scopus), the distribution of authors among gender, and career stages.

We applied the prediction model to three datasets containing the authors regarding their career development:

- Junior: researchers with a career age of fewer than five years (the first publication between 2005 and 2008)
- Mid-level: researchers with a career age between 5 and 9 years (the first publication between 2000 and 2004)

Table 1 The number of analyzed papers across scientific fields and gender and career stage distribution of authors

	Number	Percentage
Papers	40,352,318	
Health Sciences	10,608,222	26.3 %
Life Sciences	8,831,499	21.9 %
Physical Sciences	17,089,343	42.3 %
Social Sciences & Humanities	3,272,508	8.1%
Multidisciplinary	550,746	1.4%
Authors	1,824,203	
Gender:		
Female	543,517	30%
Male	1,280,686	70%
Career stage:		
Junior	265,368	15%
mid-level	533,768	29%
senior	1,025,067	56%

Table 2 Features used to train the machine learning models to predict the h-index

Feature group	Feature name	Description	Studies
Demographic	*CareerAge*	Years since first publication	[5]
	Gender	Zero for females and one for males	
	MobilityScore	Number of changing the affiliation at the country level	
	IncomeCurrentCountry	GDP Per Capita of current affiliation country	
Prior Impact	*CurrentHindex*	Current h-index	[5]; [6]; [7]
	PaperPerYear	Number of total papers divided by career age	[5]; [6]; [7]
	CitationPerPaper	Number of total citations among all papers until 2008 divided by the number of all papers	[5]; [6]; [7]
Paper/Venue	*PrimaryAuthorRatio*	Number of papers being as primary author divided by the number of all papers	
	OpenAccessRatio	Number of open access papers divided by the number of all papers	
	MainField	The scientific field with the highest amount of publications	
	HighRankPapersRatio	Number of publications in high-quality journals divided by the number of all papers	[5]
	DisciplineMobility	Number of unique disciplines authors has published paper divided by the number of all papers	
	KeywordPopularity	Number of publications with at least one popular keyword divided by the number of all papers	
	EnglishPapersRatio	Number of English papers divided by the number of all papers	
Coauthor	*MaxCoauthorHindex*	Maximum h-index of coauthors among all papers	[15]
	CoauthorPerPaper	Number of unique coauthors among all publications divided by the number of all papers	[7]
	InternationalCoauthorRatio	Number of papers with international collaboration divided by the number of all papers	

- Senior: researchers with a career age of over ten years (the first publication between 1995 and 1999).

3.2 Feature engineering

Table 2 shows variables used to estimate the future h-index of researchers. In this table, we mentioned the previous studies that employed any of the features for the prediction task. In the following, we explain how we calculated the features:

- *Gender:* It has a value equal to one for males and zero for females.
- *MobilityScore:* This feature indicates the frequency of movement between countries by tracking the authors' affiliations over their publications. More details about calculating this feature are available in our previous study [17].
- *IncomeCurrentCountry:* This feature indicates the countries' income level based on the GDP per capita of the affiliation country in the last publication. We used the World Bank information[1] to measure it.
- *PrimaryAuthorRatio:* We defined the primary author as the first or corresponding author. We computed the value of this feature by dividing the number of publications in which the researcher is the primary author to all publications.
- *OpenAccessRatio:* We extracted the article's access status from the Unpaywall dataset (a service that provides full-text articles from open access resources[2]). An open-access article can be any form of gold, green, or bronze. We declare that we could match from 8,953,939 investigated papers only 5,476,852 (61%) with Unpaywall's articles. To calculate the proportion of open access papers, we considered the number of detected as open access to the number of whole articles of the author.
- *MainField:* We identified the field of authors from the field of the journals in which they publish, and in Scopus are classified under four broad subject clusters.[3] The field with the most publications will be the main field of the author.
- *HighRankPapersRatio:* We used the journal ranking based on their quality to evaluate the rank of papers. To assess the quality of journals, we calculated the h-index of journals from 1995 to 2015. Because of different citation patterns among disciplines, journals' h-index can have varying ranges for different disciplines, which should be normalized. We applied the percentile rank approach inspired by Bornmann and Lutz [45] and computed the h-index's rank among all journals inside its discipline. We used Scopus's classification system to find the journals' disciplines. In this system, journals are classified into 27 subject categories.[4] In this percentile rank approach, each journal within a category ranks 0 (lowest h-index) to 100 (highest h-index). Journals with the same h-index have the same rank. If the journal belongs to more than one category, we used the weighted Percentile Ranking wPR) [46]. Based on this approach, wPR will be calculated using the formula:

$$wPR = \frac{PR_{sc1} * n_{sc1} + PR_{sc2} * n_{sc2} + \cdots + PR_{sci} * n_{sci}}{n_s c1 + n_s c2 + \cdots + n_{sci}}. \quad (1)$$

Whereby *sci* is the *i*th subject category that the journal belongs to and n_{sci} is the number of journals in this subject category, and PR_{sci} is PR of the journal in it. Journals with a wPR higher than 50% are assumed to be high quality. Finally, we counted the proportion of the author's publications in high-quality journals among all their publications for the variable *HighRankPapersRatio*.

[1] https://www.weforum.org/agenda/2020/08/world-bank-2020-classifications-low-high-income-countries/.

[2] https://unpaywall.org/.

[3] https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/scopus/~/what-are-the-most-frequent-subject-area-categories-and-classifications-used-in/.

[4] https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/scopus/~/what-are-the-most-frequent-subject-area-categories-and-classifications-used-in/.

Table 3 Descriptive statistics of features. This table shows the mean standard deviation for numerical features and distribution of authors based on their gender, mobility status and main field

Feature name	Mean	Standard deviation	Distribution
CareerAge	9.35	3.69	
Gender	0.70	0.46	70% male, 30% female
MobilityScore	0.50	1.08	27% mobile, 73% non-mobile
IncomeCurrentCountry	35,052.63	14,024.40	
CurrentHindex	6.13	6.17	
PaperPerYear	2.00	2.39	
CitationPerPaper	11.47	22.18	
PrimaryAuthorRatio	0.36	0.29	
OpenAccessRatio	0.19	0.23	
MainField			H: 29%, L:23%, P:37%, S:6%, M:4% *
HighRankPapersRatio	0.01	0.06	
DisciplineMobility	0.47	0.45	
KeywordPopularity	0.53	0.28	
EnglishPapersRatio	0.92	0.20	
MaxCoauthorHindex	15.51	14.86	
CoauthorPerPaper	3.74	30.39	
InternationalCoauthorRatio	0.21	0.25	

*H: Health Sciences, L: Life Sciences, P: Physical Sciences, M:Multiple Fields.

- *DisciplineMobility:* This feature indicates the number of unique fields the author has published during the entire academic age divided by the number of whole papers.
- *KeywordPopularity:* This feature indicates the proportion of papers with popular keywords. First, we ranked keywords based on the frequency of occurrence in papers from the same discipline (27 subject categories) and publication year to measure the keyword popularity for a paper. Next, we gave a value of 1 to the paper with a ranking above 0.5; otherwise, 0. Finally, we summed up these values over all papers and divided them by the number of all papers.
- *EnglishPapersRatio:* This feature measures the ratio of papers written in English.
- *CoauthorPerPaper:* This feature displays the number of unique coauthors, which is normalized by dividing by the number of all papers.
- *CoauthorMaxHindex:* To assess the effect of the scientific impact of coauthors, we used the maximum h-index among all coauthors as an alternative measure of the Godfather Effect [15].
- *InternationalCoauthorRatio:* This feature specifies the number of international collaborators for all papers. To calculate it, first, we counted the number of papers with at least one coauthor having a different country in the affiliation than the author and then divided it by the number of all papers.

We provided descriptive statistics for investigated features in Table 3 to describe the data.

3.3 Applied methods for the prediction task

We tackled the h-index prediction as a regression problem comparable to previous studies [5–7, 11, 16]. We explored the performance of four different machine learning methods, namely SVR, RF, GB, and XGBoost. Among these, XGBoost emerged as the top-performing method, consistent with the findings reported by [7]. Consequently, we utilized the XGBoost approach for our h-index prediction task. XGBoost is a scalable end-to-end tree boosting system introduced by Chen and Guestrin [47]. It efficiently implements

Table 4 Different feature combinations to predict the h-index

Feature group	Feature name	Feature combination								
		1	2	3	4	5	6	7	8	9
Demographic	*CareerAge*		✓		✓			✓		✓
	Gender		✓		✓			✓		✓
	MobilityScore		✓		✓			✓		✓
	IncomeCurrentCountry		✓		✓			✓		✓
Prior impact	*CurrentHindex*	✓	✓	✓	✓	✓				
	CitationPerPaper	✓	✓	✓	✓	✓				
Paper/venue	*PrimaryAuthorRatio*	✓	✓	✓	✓		✓	✓	✓	✓
	OpenAccessRatio	✓	✓	✓	✓		✓	✓	✓	✓
	MainField	✓	✓	✓	✓		✓	✓	✓	✓
	HighRankPapersRatio	✓	✓	✓	✓		✓	✓	✓	✓
	DisciplineMobility	✓	✓	✓	✓		✓	✓	✓	✓
	EnglishPapersRatio	✓	✓	✓	✓		✓	✓	✓	✓
	KeywordPopularity	✓	✓	✓	✓		✓	✓	✓	✓
Coauthor	*MaxCoauthorHindex*			✓	✓				✓	✓
	CoauthorPerPaper			✓	✓				✓	✓
	InternationalCoauthorRatio			✓	✓				✓	✓

Gradient Boosting in terms of speed and is appropriate for solving problems using minimal resources. We need to have the data in numerical form to apply this method. We utilized one hot encoder to convert the categorical values to integers. In this encoding method, each value of the categorical variable will be converted to a feature with a binary value, where 1 represents the data value and 0 is used for all other values. So, for *MainField* with five values, we have five features, and the feature with a value equal to 1 indicates the *MainField*. To evaluate the model, we utilized the Mean Absolute Percentage Error (MAPE) to measure the error as a percentage, which is appropriate to compare the performance of a model for the different datasets, as used by some previous studies [6–8]. Because MAPE is affected by outliers [48], we also utilized symmetric Mean Absolute Percentage Error (sMAPE), which is scaled to percentage too and is more resistant to outliers [47]. In addition, we used Root Mean Square Error (RMSE) to evaluate the performance of models, as in prior works [5, 8, 9]. We used the 5-fold cross-validation procedure to evaluate the models.

We defined different feature combinations based on the attributes of the author, paper, venue, and coauthors to see which feature categories are better for short/long-term prediction. Table 4 shows the different feature combinations utilized to train the model.

Prior studies regarded varying time frames to estimate the future h-index [5, 7, 49] and examined several years from one to five-year and [49] for five-year and ten-year time frames. The prediction performance declined as the prediction time frame increased in all studies. We considered the h-index as our target from one to ten years in the future (h-index from 2009 to 2018). It enables us to measure the extent of predicting performance in the future.

To examine the importance of each feature in the prediction task, we employed a feature selection technique, *Recursive Feature Elimination* (RFE), which removes recursively features and builds a model based on the remaining features [50, 51].

4 Results

In this section, we present the results of our analysis, focusing on the relationship between various features and the future h-index of researchers. Before delving into the specific

findings, we address the potential multicollinearity problem in Sect. 4.1 by examining the dependencies between features. We analyze the Pearson correlation between independent variables and visualize the results using a heatmap. Next, we explore the correlation between the introduced features and the future h-index in 2009, 2014, and 2018. This analysis allows us to examine the statistical association between variables, providing insights into the strength and direction of these relationships. However, it's important to note that the correlations captured by the correlation analysis primarily represent linear associations between features and the h-index.

To capture the non-linear relationship between the h-index and the investigated features, we apply ML prediction models in Sect. 4.2. First, in Sect. 4.2.1, we identify the most important factors for predicting the h-index using the feature selection method, RFE. This step helps us narrow down the key variables. Then, in Sect. 4.2.2, we examine the effectiveness of these models for researchers with different career ages, focusing on the temporal dimension.

4.1 Correlation analysis

Before investigating the relationship between various features and future h-index, we examine the dependencies between features to avoid the potential multicollinearity problem. Figure 1 presents the Pearson correlation between independent variables. We see a strong correlation between *PaperPerYear* and *CurrentHindex*; therefore, to avoid multicollinearity in regression and classification models, we exclude *PaperPerYear* from the data for prediction tasks.

To examine the affecting factors on the h-index, we first provide the correlation between features introduced in Table 2 and future h-index. Table 5 presents the Pearson correlation coefficient between the features (except for *MainField*, a categorical variable) and h-index in 2009, 2014, and 2018. The highest correlation coefficient for two prior impact-based features (*CurrentHindex*, *PaperPerYear*) displays the strong association of this kind of feature with the future h-index. The higher correlation coefficient between the future h-index and the number of papers (*PaperPerYear*) than the number of citations (*CitationPerPaper*) reveals that productivity has a more significant impact than received citations on the h-index. Among non-prior impact-based features, *MaxCoauthorHindex* has the highest correlation with the h-index and suggests the strong relation of coauthors' reputation with the future h-index. The negative value for *DisciplineMobility* suggests that authors who publish in several scientific fields have a lower h-index than those who publish in a specific field.

Most of the correlations between the influential factors and the h-index demonstrate consistent patterns across different time frames, indicating similar effects in both the short and long term. While correlation analysis offers informative perspectives about the strength and direction of these relationships, it primarily captures linear associations between variables. However, we will employ machine learning algorithms in the next section to uncover non-linear associations and delve deeper into the temporal dimension of the relationship for researchers in different career stages. This approach allows us to examine the complex interactions and temporal dynamics between the factors and the h-index, specifically analyzing how they vary across different career stages. It provides a more comprehensive understanding of their relationship and enables us to make accurate predictions beyond what correlation analysis alone can reveal.

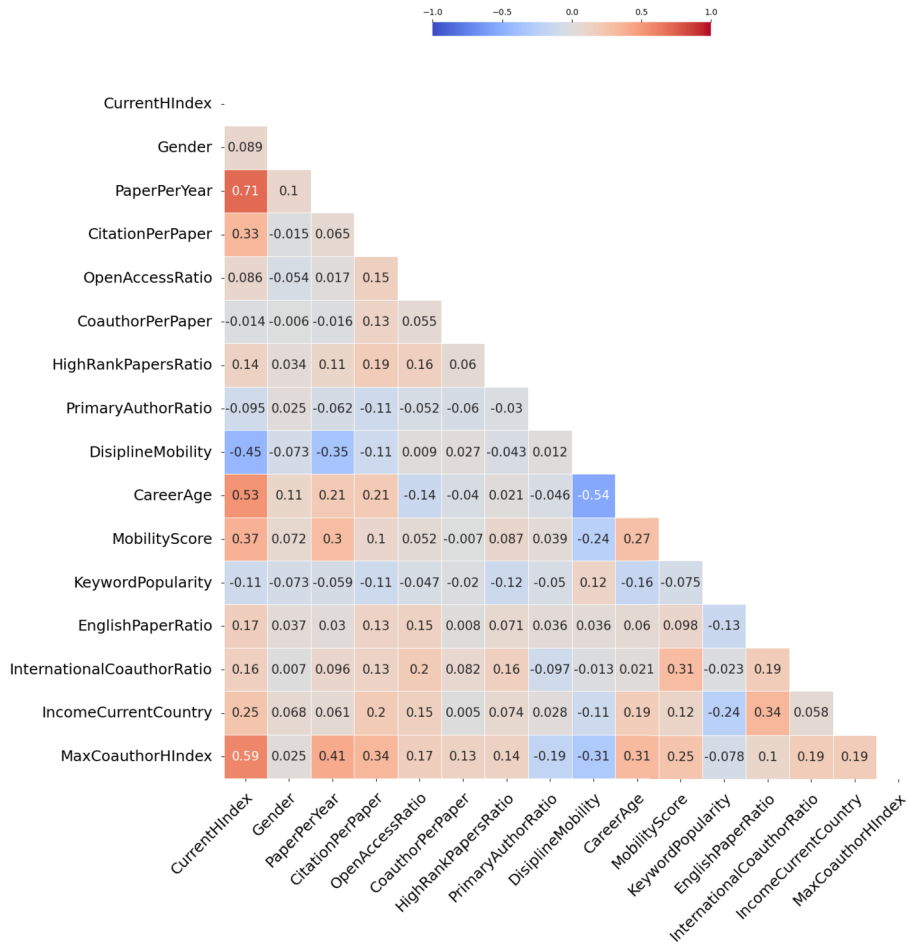

Figure 1 Pearson correlation heatmap of independent variables. We observe a particularly strong positive correlation between 'PaperPerYear' and 'CurrentHindex' in this heatmap

Table 5 Pearson correlation coefficient between the features and h-index in the future for three different years. *CurrentHindex*, *PaperPerYear*, and *CitationPerPaper* are prior impact-based features, and the rest are non-prior impact-based features

Feature	H-index		
	2009	2014	2018
CareerAge	0.48	0.38	0.32
Gender	0.09	0.08	0.07
MobilityScore	0.44	0.43	0.41
IncomeCurrentCountry	0.23	0.21	0.19
CurrentHindex	0.99	0.95	0.87
PaperPerYear	0.73	0.75	0.73
CitationPerPaper	0.31	0.26	0.23
PrimaryAuthorRatio	−0.09	−0.08	0.-0.06
OpenAccessRatio	0.10	0.14	0.15
EnglishPapersRatio	0.17	0.16	0.15
KeywordPopularity	−0.09	−0.07	−0.05
HighRankPapersRatio	0.14	0.15	0.15
DisciplineMobility	−0.45	−0.42	−0.39
MaxCoauthorHindex	0.58	0.58	0.55
CoauthorPerPaper	−0.01	0.02	0.04
InternationalCoauthorRatio	0.17	0.19	0.19

4.2 Prediction analysis

In this section, we present the prediction results of our study, highlighting the influence of different features on predicting the h-index. Firstly, in Sect. 4.2.1, we evaluate the importance of these features using the Recursive Feature Elimination (RFE) method. Then, in Sect. 4.2.2, we examine the effectiveness and stability of various feature combinations in predicting the h-index. We analyze the predictive performance across different time frames and for researchers at different career stages, providing valuable insights into the temporal dynamics and the impact of features on the h-index prediction task.

4.2.1 Feature impact

We evaluate the importance of features in the prediction task by ranking them via the RFE method. Table 6 demonstrates the feature ranking for selecting the predictors in the model. For *MainField*, we used one hot encoder, which converts each unique category value to a feature (five features for five fields). The features highlighted in blue are the top five features in the selection process. We observe that paper-specific features are most relevant among all career stages. Also, coauthor-specific features are among the most important features to predict the h-index for the researchers in junior and mid-level career stages. It suggests that the coauthor's characteristics have more influence on the h-index for these researchers than seniors.

4.2.2 Career stage and temporal dimension of model performance

Before we show the result of the analyses, we make some comparisons between the performance of our model and previous works. Wu et al. [7] have already compared their performance with other studies [5, 6, 49] and presented the best performance among all

Table 6 Ranking of features for selection in predicting the h-index with the RFE method. The five most relevant features (with a ranking between 1 and 5) are highlighted in blue. It demonstrates variations in feature importance across career stages and prediction years. 'CurrentHindex' consistently ranks as the top feature, indicating its significant influence. Additionally, the most influential features vary by career stage, highlighting the complexity of research impact factors

Career stage	**Junior**			**Mid-level**			**Senior**		
Prediction year	**2009**	**2014**	**2018**	**2009**	**2014**	**2018**	**2009**	**2014**	**2018**
Feature:	Rank	Rank	Rank	Rank	Rank	Rank	Rank	Rank	Rank
CareerAge	7	5	4	3	2	3	3	4	2
Gender	20	18	16	19	18	16	19	19	19
MobilityScore	18	14	12	12	8	4	18	18	16
IncomeCurrentCountry	14	16	17	13	14	13	9	9	9
CurrentHindex	1	1	1	1	1	1	1	1	1
CitationPerPaper	11	15	15	6	6	7	7	5	6
PrimaryAuthorRatio	6	10	9	10	9	9	16	11	10
OpenAccessRatio	3	8	7	4	7	8	6	2	5
EnglishPapersRatio	4	11	13	9	16	17	15	17	18
KeywordPopularity	10	13	14	11	13	15	11	14	13
MainField									
Health Sciences	12	3	5	17	19	18	5	15	17
LifeSciences	15	17	18	15	17	19	8	6	3
multiple fields	19	20	20	20	20	20	20	20	20
Physical Sciences	13	2	2	16	4	5	13	7	8
Social Sciences	16	19	19	14	15	14	4	3	4
HighRankPapersRatio	9	9	11	5	12	12	12	16	15
DisciplineMobility	2	12	10	2	11	11	2	12	14
MaxCoauthorHindex	5	6	3	8	5	6	10	10	11
CoauthorPerPaper	17	4	6	18	10	10	17	8	7
InternationalCoauthorRatio	8	7	8	7	3	2	14	13	12

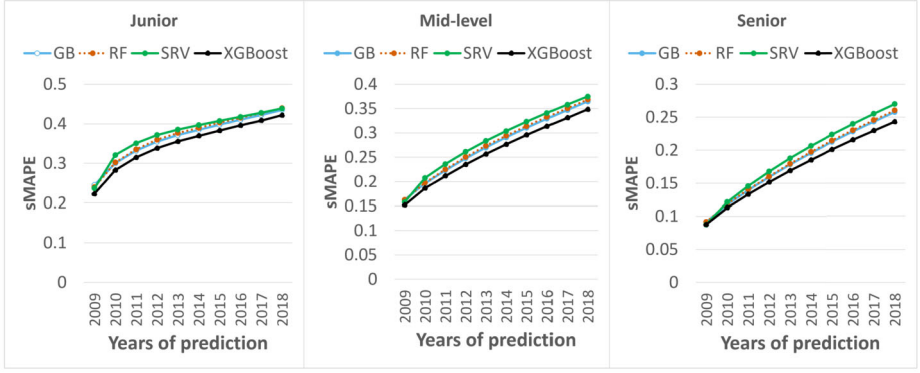

Figure 2 Comparison of predictive performance using sMAPE metric among four machine learning algorithms (SVR, RF, GB, and XGBoost) for researchers' h-index prediction at different career stages from 2009 to 2018. The analysis utilized feature combination 1 as the predictor

these studies. They excluded the authors with an h-index of less than four from the investigated data. They achieved the minimum MAPE of 0.063 for the first prediction year by employing more prior impact-based features. We could reach the minimum MAPE of 0.068 by applying this condition to investigated authors. Instead, two-thirds of the authors will be discarded in our analyses. Because of losing too much data, particularly from young scholars, we didn't apply this condition and implemented our models with all authors, despite reducing the performance. To evaluate the predictive performance, we conducted a comparison among four machine learning algorithms: SVR, RF, GB, and XGBoost, using feature combination 1, which includes all features. The results are illustrated in Fig. 2, demonstrating that XGBoost outperforms the other methods across all career stages. As a result, we proceed with this method for further analyses.

Table 7 showcases the performance metrics, including RMSE, MAPE, and sMAPE, for all three groups of researchers (junior, middle-level, and senior) across the years 2009, 2014, and 2018. It provides a detailed overview of the model's performance, enabling a direct comparison of the metrics for each group and year. Lower values of these metrics indicate better predictive performance. We observe a decline in performance for all groups of researchers across all metrics from the near future (2009) to the far future (2018). While the models for seniors generally demonstrate better performance compared to the other groups, the decline in performance is more pronounced for researchers in later career stages. Specifically, in terms of RMSE for junior researchers, the range varies from 0.6 (combination 4, considering all features) in 2009 to 5.46 (combination 1, considering only prior-impact features) in 2018. For seniors, the range is from 0.74 (combination 1) in 2009 to 6.93 (combination 8) in 2018. We observe a greater decline in performance for seniors in the far future compared to juniors. When considering MAPE and sMAPE, which provide performance in percentage, we can better compare the model's performance across career stages. Although these metrics show better performance for researchers in later career stages, the performance is more stable for juniors. For instance, combination 4 exhibits the best performance for juniors, with sMAPE ranging from 0.22 to 0.42, while for seniors, it ranges from 0.09 to 0.24. Furthermore, despite combinations containing prior-impact features exhibiting better performance in the near future (2009) for all researcher groups, we observe that for juniors, combinations without prior-impact features approach

Table 7 Comparison of XGBoost regression model performance to predict the feature h-index in one, five, and ten years (2009, 2014, and 2018) implemented on three datasets (junior, middle, and senior researchers). RMSE, MAPE, and sMAPE are the metrics to assess performance

Feature combination	Metric	Junior			Middle-level			Senior		
		2009	2014	2018	2009	2014	2018	2009	2014	2018
1	RMSE	0.62	3.01	5.15	0.68	2.85	4.94	0.75	3	5.09
	MAPE	0.24	0.52	0.62	0.16	0.33	0.45	0.09	0.2	0.28
	sMAPE	0.23	0.39	0.45	0.16	0.29	0.36	0.09	0.19	0.25
2	RMSE	0.61	2.91	4.99	0.67	2.78	4.81	0.75	2.94	4.97
	MAPE	0.24	0.49	0.59	0.16	0.32	0.43	0.09	0.2	0.28
	sMAPE	0.23	0.38	0.43	0.15	0.28	0.35	0.09	0.19	0.25
3	RMSE	0.61	2.85	4.91	0.68	2.75	4.77	0.75	2.9	4.9
	MAPE	0.24	0.5	0.6	0.16	0.33	0.44	0.09	0.2	0.28
	sMAPE	0.23	0.38	0.44	0.15	0.28	0.36	0.09	0.19	0.25
4	RMSE	0.6	2.78	4.81	0.67	2.68	4.67	0.74	2.85	4.8
	MAPE	0.24	0.48	0.57	0.16	0.32	0.43	0.09	0.2	0.27
	sMAPE	0.22	0.37	0.42	0.15	0.28	0.35	0.09	0.19	0.24
5	RMSE	0.67	3.23	5.46	0.72	3.05	5.23	0.78	3.24	5.49
	MAPE	0.28	0.57	0.68	0.17	0.37	0.49	0.09	0.23	0.31
	sMAPE	0.27	0.42	0.47	0.17	0.31	0.39	0.1	0.21	0.28
6	RMSE	1	3.27	5.43	1.87	3.56	5.5	4.04	5.75	7.52
	MAPE	0.37	0.56	0.65	0.41	0.44	0.53	0.41	0.4	0.44
	sMAPE	0.31	0.41	0.46	0.32	0.35	0.4	0.32	0.32	0.34
7	RMSE	0.97	3.19	5.3	1.8	3.48	5.38	3.79	5.47	7.24
	MAPE	0.36	0.54	0.62	0.39	0.43	0.51	0.38	0.38	0.42
	sMAPE	0.31	0.4	0.44	0.31	0.34	0.39	0.31	0.31	0.33
8	RMSE	0.96	2.97	5.02	1.75	3.33	5.23	3.64	5.23	6.93
	MAPE	0.35	0.53	0.62	0.38	0.41	0.5	0.35	0.36	0.4
	sMAPE	0.3	0.4	0.44	0.31	0.33	0.38	0.29	0.29	0.32
9	RMSE	0.94	2.92	4.93	1.69	3.28	5.15	3.47	5.05	6.74
	MAPE	0.34	0.51	0.6	0.36	0.41	0.49	0.34	0.34	0.39
	sMAPE	0.3	0.39	0.43	0.3	0.33	0.38	0.28	0.29	0.32

the performance of models with prior-impact features in the long term (2018). In some cases, these combinations even outperform models with prior-impact features. This finding suggests that non-prior impact-based features are more reliable predictors for the future h-index of junior researchers, compared to seniors. In summary, seniors generally exhibit better performance, but juniors demonstrate more stable performance and the potential for improved long-term predictions using non-prior impact-based features.

To further illustrate the performance trends over time, Fig. 3 focuses on the sMAPE metric and covers the years from 2009 to 2018. It offers a visual representation of the prediction efficiency of different feature combinations for researchers at different career stages throughout the entire time span. In this figure, the lower sMAPE for combinations including prior impact-based features indicates the higher performance for these combinations, but losing the performance with the passing years for these combinations is more than other combinations.

To compare the prediction efficiency between different career stages, we implemented the prediction model for authors from three career stages and presented the performance (sMAPE) in Fig. 3(a). We observe a better performance for the combinations containing prior impact-based features for all researchers' groups in the near future. Still, they lose more performance than combinations without prior impact-based features in the distant future. Interestingly, the performance of non-prior impact-based models (e.g., combina-

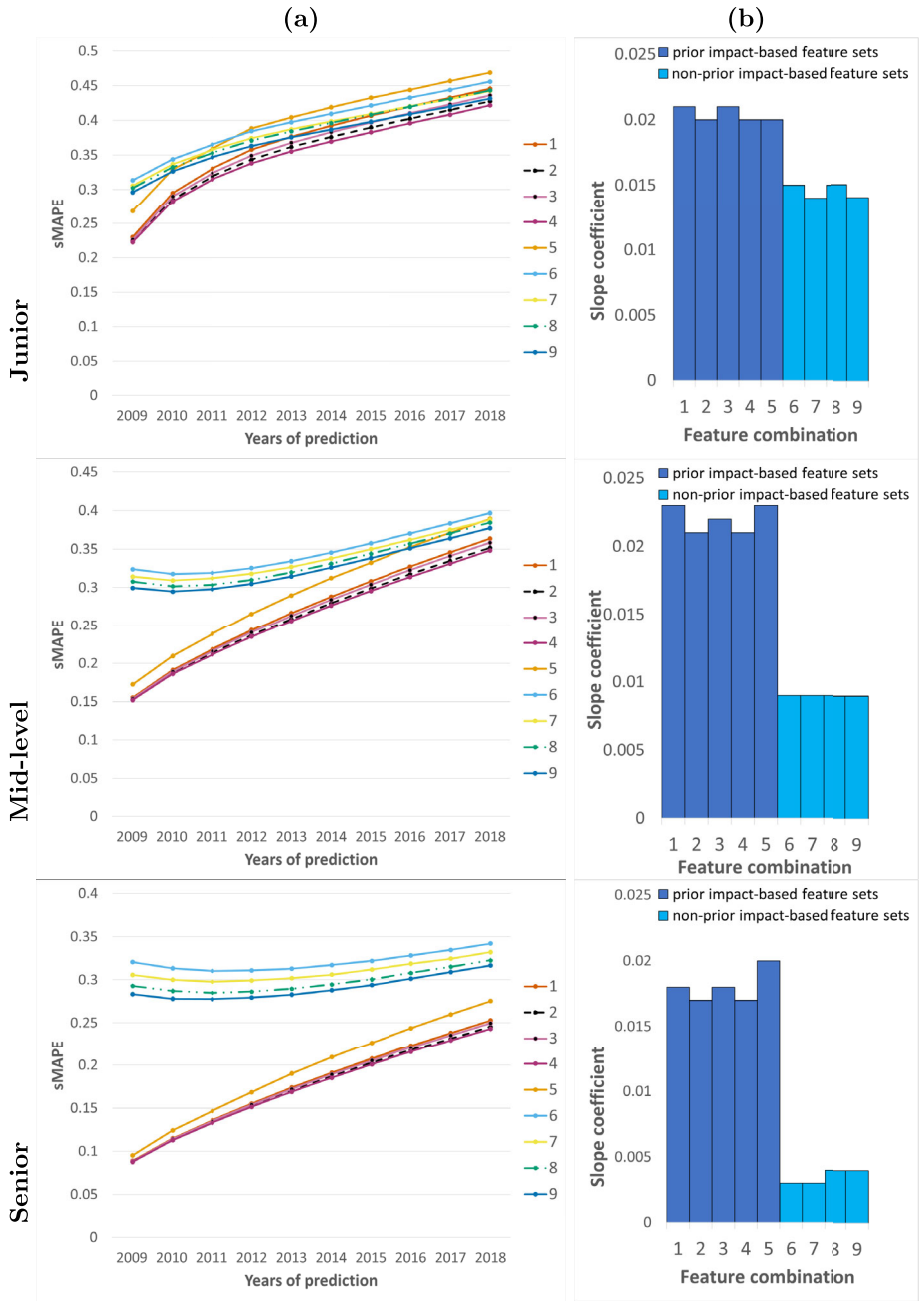

Figure 3 Comparison of predictive performance (a) and slope coefficients (b) over ten years for different feature combinations trained with the XGBoost regression method among researchers of varying experience levels (junior, mid-level, and senior). (a) illustrates the performance of predicting models using the sMAPE metric. (b) displays the corresponding slope coefficients, indicating the performance change over time. The dark/light blue columns in (b) represent feature combinations, including/excluding prior impact-based features

tions 8 and 9) for junior researchers, which is worse than prior impact-based models (e.g., combinations 1 and 5) in the earlier years, dominates them in the long term. We see a similar result for researchers at the mid-level (better performance for combinations 8 and 9

than combination 5). This suggests that non-prior impact-based features are more reliable in predicting the future h-index of younger researchers over distant periods.

To quantify the extent of performance degradation for the two groups of combinations (prior and non-prior impact-based features), we calculated the slope coefficient for model performances reported in Fig. 3(a). The slope coefficient (m) was computed using the least squares method [52] with the following equation:

$$m = \frac{\sum(x - \bar{x}y - \bar{y})}{\sum(x - \bar{x})^2}, \tag{2}$$

where x represents the years from 2009 to 2018, y represents the sMAPE in the corresponding year and $\bar{x}$ and $\bar{y}$ are their respective averages over the ten-year period.

The presented slope coefficient in Fig. 3(b) reveals insights into the stability of the models' performance. A lower slope coefficient signifies greater stability, indicating that the model's performance changes more slowly and consistently over the ten-year period. Conversely, a higher slope coefficient indicates that the model's performance fluctuates more significantly.

In general, we observed a higher slope coefficient (indicating more significant performance loss over time) for feature combinations with prior impact-based features (in dark blue) compared to other feature combinations for researchers at any career stage. The lower value for combinations containing non-prior impact-based features (in light blue) indicates that they are more stable predictors in the long term, although at a modest performance level.

5 Limitations

In this study, we considered just journal papers and not conference papers, and it causes bias issues, especially for disciplines in which authors publish their studies mainly as conference proceedings papers. Another limitation is the problem concerning data reliability and validity in calculating the features. For example, to obtain the proportion of open-access publications, we identified the access form of articles in 2019 on Unpaywall. Many journals have changed their business model to open-access or closed-access. We can not be sure about the accessibility of papers at the time of publishing and two years time windows that we considered to calculate the number of received citations. Also, we measured the mobility feature similar to our previous paper [17], and the mentioned limitations in that paper exist for this feature too.

6 Main findings and discussion

In this study, we comprehensively investigated the impact of different feature categories on predicting the h-index for researchers at various career stages. By employing a machine learning approach and extensive feature analysis, our main objective was to understand the factors influencing researchers' future scholarly impact and how these factors differ based on their career stage.

The contributions of this research are threefold, as outlined in the introduction. Firstly, we explored the impact of various features on predicting researchers' h-index across different career stages by employing the feature selection technique, RFE, and implementing predictive models for various feature combinations. This analysis gave us valuable insights into the predictive power of different attributes and their varying effectiveness at different

career phases. Our analysis of Table 7 and Fig. 3(a) revealed that models with prior impact-based features demonstrated better performance than those without these features. This finding suggests that prior impact-based features are more reliable predictors of future scholarly impact, particularly for researchers in later career stages, both in the short and long term. Conversely, the smaller performance gap between models with prior impact-based feature combinations and models without such features for junior researchers in the short term, and the superiority of models with non-prior impact-based features over models with prior impact-based features in the long term (as shown in Table 7), indicates that non-prior impact-based features play a more prominent role, particularly in long-term predictions, for younger researchers. This implies that these non-prior impact-based features could be valuable for identifying rising stars with strong potential for future scientific impact.

Secondly, our investigation delved into the temporal dimension of feature performance, encompassing both prior impact-based and non-prior impact-based features. We made notable observations by examining different feature combinations and their predictive power over time. Prior impact-based features exhibited the highest predictive accuracy in the short term, but their performance significantly declined in the long term compared to other features. This finding underscores the importance of considering non-prior impact-based features for enhancing long-term predictions.

Lastly, we introduced novel author (e.g., demographic characteristics) and paper/venue-specific features to estimate the author's h-index and assessed their impact on prediction tasks through feature selection analysis. The results revealed interesting insights into the individual contributions of these features to researchers' scientific impact. Among the introduced features, gender showed the weakest predictive power, suggesting that gender has almost no impact on the scientific impact, which is desirable. However, *OpenAccessRatio* emerged as one of the top five powerful predictors for junior and mid-level seniors in the short term and held a similar position for seniors in the long term. In contrast, *DisciplineMobility* ranked as the second top predictor for researchers from any career stage in the short term but exhibited weaker predictive power in the long term. The higher ranking of *MaxCoauthorHindex* in predicting the h-index for researchers in earlier career stages, both in the short and long term, highlighted the significance of co-authors and their reputation in forecasting future h-index values. Additionally, *InternationalCoauthorRatio* was among the top five predictors for mid-level researchers in the long term, while the *MainField* also held a place among the top five predictors, indicating a strong association of the h-index with specific research fields. Notably, *SocialSciences* featured as one of the top predictors for senior researchers, while *PhysicalSciences* played a similar role for junior and mid-level researchers in the long term, suggesting that predicting the h-index of seniors and certain disciplines in the long term is more feasible. On the other hand, *MobilityScore* demonstrated no significant impact on the h-index for any of the three groups of researchers, except for mid-level researchers in the long term, where it ranked fourth. Finally, other newly introduced features, such as *KeywordPopularity* and *PrimaryAuthorRatio*, had minimal impact due to their low ranking in the feature selection process.

Additionally, the results of the correlation analysis were consistent with the feature selection findings. A positive moderate correlation coefficient was observed between the authors' international mobility and their future h-index. However, given the low proportion of mobile researchers (about 27%), this author's feature proved less effective in predicting

the h-index when accounting for other factors. Conversely, we found a very weak correlation between gender and the h-index, with gender displaying the lowest importance in predicting the h-index among all features. The results also underscored the importance of focusing on the study's field to achieve a better scientific impact. Paper/venue-specific features were shown to have more impact on the future h-index than the author's demographic and co-authorship characteristics.

The performances of proposed models indicate that still more features that don't depend on the history of publications and citations are required to forecast the future h-index of young researchers. For example, [13, 15] focused on analyzing the co-authorship network to investigate the relationship between the structural role of authors in the network and the future h-index. Using such intensive network analysis in our study could improve the performance, particularly for junior researchers with lower impact history in their profiles. Additionally, the textual content of papers examined by [13] and topic authority by [49] could be combined with the introduced features in this study to enhance the predictive power of our models. By incorporating these additional features alongside the ones introduced in our research, we may offer a more comprehensive understanding of researchers' future scholarly impact and lead to more accurate predictions for early-career academics.

7 Conclusion

This study aims to reveal the factors associated with the future h-index of researchers based on bibliometric data, which allowed us to have various researchers groups from different countries and scientific fields for more comprehensive analyses. The results can be informative for researchers to understand how bibliometric characteristics of authors and papers can influence the future h-index and for policymakers to support them by focusing on the factors having positive relations with scientific success. We admit that the h-index, which is the most popular metric to assess the scholars, suffers from some limitations (e.g., field-dependent [53], incapable of comparing researchers in different career stages [24] and detect authors with extremely highly cited papers [54], can be manipulated by self-citations [55]). Our work is not about promoting the h-index, but acknowledging its deficiencies to better understand what factors influence it. Without understanding these factors, researchers cannot understand its biases. Hence we actually contribute to understanding the deficiencies. In addition, possible bias by missing data (e.g., including only authors with gender status) can affect the validity of models. In addition, margin error has not been indicated in this study, and the reliability level of these models is uncertain.

To predict the scientific impact, we employed artificial intelligence (AI) models, which are supposed to mimic human decision-making for assessment and don't necessarily lead to ethical and desirable results. One ethical issue is considering certain features that cause discriminatory effects or introduce bias against certain groups in the predicting model [56, 57], which we don't intend in this study. For example, investigating gender as a predictor in the prediction model was to study gender inequality in science for more attention in policy-making.

Acknowledgements
We acknowledge the support of the German Competence Center for Bibliometrics (grant: 01PQ17001) for maintaining the used dataset for the analyses.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work is financially supported by BMBF project OASE, grant number 01PU17005A.

Abbreviations
GB, Gradient Boosting; GBDT, Gradient-Boosting Decision Tree; GBRT, Gradient Boosted Regression Trees; GDP, Gross Domestic Product; KNN, K-nearest neighbour; MAPE, Mean Absolute Percentage Error; NN, Neural Networks; OACA, Open Access Citation Advantage; RF, Random Forest; RFE, Recursive Feature Elimination; RMSE, Root Mean Square Error; sMAPE, symmetric Mean Absolute Percentage Error; SVR, Support Vector Regression; wPR, weighted Percentile Ranking; XGBoost, Extreme Gradient Boosting.

Availability of data and materials
We don't have permission to redistribute Spocus's raw data, but processed data used for the analyses are available and documented in the Git repository Git repository.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contributions
SD supervised this study, and PM was the project leader that supported it financially. Material preparation, data collection, Methodology, analysis, Validation, and Visualization were performed by FM. FM wrote the first draft of the manuscript, and all authors commented on previous versions. All authors read and approved the final manuscript.

Authors' information
Fakhri Momeni is a research associate at GESIS – Leibniz Institute for the Social Sciences in Cologne and Ph.D. student in information science at Heinrich Heine University in Duesseldorf. Dr. Philipp Mayr is a team leader (Information & Data Retrieval) at GESIS in Cologne, department Knowledge Technologies for the Social Sciences (KTS). Prof. Dr. Stefan Dietze is Professor of Data & Knowledge Engineering at Heinrich Heine University Duesseldorf and Scientific Director of the Knowledge Technologies department for the Social Sciences at GESIS in Cologne.

Author details
[1]GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany.
[2]Heinrich-Heine-University, Universitätsstr. 1, 40225 Düsseldorf, Germany.

Received: 8 July 2022 Accepted: 23 September 2023 Published online: 06 October 2023

References

1. Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci 102(46):16569–16572
2. Egghe L et al (2006) An improvement of the h-index: the g-index. ISSI Newsl 2(1):8–9
3. Kaur J, Radicchi F, Menczer F (2013) Universality of scholarly impact metrics. J Informetr 7(4):924–932
4. Daud A, Abbasi R, Muhammad F (2013) Finding rising stars in social networks. In: International conference on database systems for advanced applications. Springer, Berlin, pp 13–24
5. Ayaz S, Masood N, Islam MA (2018) Predicting scientific impact based on h-index. Scientometrics 114(3):993–1010
6. Weihs L, Etzioni O (2017) Learning to predict citation-based impact measures. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, Los Alamitos, pp 1–10
7. Wu Z, Lin W, Liu P, Chen J, Mao L (2019) Predicting long-term scientific impact based on multi-field feature extraction. IEEE Access 7:51759–51770
8. Bai X, Zhang F, Lee I (2019) Predicting the citations of scholarly paper. J Informetr 13(1):407–418
9. Abrishami A, Aliakbary S (2019) Predicting citation counts based on deep neural network learning techniques. J Informetr 13(2):485–499
10. Jiang S, Koch B, Sun Y (2021) Hints: citation time series prediction for new publications via dynamic heterogeneous information network embedding. In: Proceedings of the web conference 2021, pp 3158–3167
11. Ruan X, Zhu Y, Li J, Cheng Y (2020) Predicting the citation counts of individual papers via a bp neural network. J Informetr 14(3):101039
12. Kossmeier M, Heinze G (2019) Predicting future citation counts of scientific manuscripts submitted for publication: a cohort study in transplantology. Transpl Int 32(1):6–15
13. Nikolentzos G, Panagopoulos G, Evdaimon I, Vazirgiannis M (2021) Can author collaboration reveal impact? The case of h-index pp 177–194
14. 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
15. McCarty C, Jawitz JW, Hopkins A, Goldman A (2013) Predicting author h-index using characteristics of the co-author network. Scientometrics 96(2):467–483
16. Dong Y, Johnson RA, Chawla NV (2016) Can scientific impact be predicted? IEEE Trans Big Data 2(1):18–30
17. Momeni F, Karimi F, Mayr P, Peters I, Dietze S (2022) The many facets of academic mobility and its impact on scholars' career. J Informetr 16(2):101280
18. Singh V (2018) Comparing research productivity of returnee-phds in science, engineering, and the social sciences. Scientometrics 115(3):1241–1252
19. Netz N, Hampel S, Aman V (2020) What effects does international mobility have on scientists' careers? A systematic review. Res Eval 29(3):327–351
20. Liu J, Wang R, Xu S (2021) What academic mobility configurations contribute to high performance: an fsqca analysis of csc-funded visiting scholars. Scientometrics 126(2):1079–1100

21. Radford DM, Parangi S, Tu C, Silver JK (2022) h-index and academic rank by gender among breast surgery fellowship faculty. J Women's Health 31(1):110–116
22. Carter TE, Smith TE, Osteen PJ (2017) Gender comparisons of social work faculty using h-index scores. Scientometrics 111(3):1547–1557
23. Lopez SA, Svider PF, Misra P, Bhagat N, Langer PD, Eloy JA (2014) Gender differences in promotion and scholarly impact: an analysis of 1460 academic ophthalmologists. J Surg Educ 71(6):851–859
24. Kelly CD, Jennions MD (2006) The h index and career assessment by numbers. Trends Ecol Evol 21(4):167–170
25. Leydesdorff L, Bornmann L, Wagner CS (2019) The relative influences of government funding and international collaboration on citation impact. J Assoc Inf Sci Technol 70(2):198–201
26. Smirnova N, Mayr P (2023) A comprehensive analysis of acknowledgement texts in web of science: a case study on four scientific domains. Scientometrics 128(1):709–734
27. Gantman ER (2012) Economic, linguistic, and political factors in the scientific productivity of countries. Scientometrics 93(3):967–985
28. Confraria H, Godinho MM, Wang L (2017) Determinants of citation impact: a comparative analysis of the global south versus the global North. Res Policy 46(1):265–279
29. Malesios C, Psarakis S (2014) Comparison of the h-index for different fields of research using bootstrap methodology. Qual Quant 48(1):521–545
30. Lillquist E, Green S (2010) The discipline dependence of citation statistics. Scientometrics 84(3):749–762
31. Iglesias J, Pecharromán C (2007) Scaling the h-index for different scientific isi fields. Scientometrics 73(3):303–320
32. Petersen AM, Penner O (2014) Inequality and cumulative advantage in science careers: a case study of high-impact journals. EPJ Data Sci 3:1
33. Xie F, Ghozy S, Kallmes DF, Lehman JS (2022) Do open-access dermatology articles have higher citation counts than those with subscription-based access? PLoS ONE 17(12):0279265
34. Blair LD, Odell JD (2020) The open access policy citation advantage for a medical school
35. Ottaviani J (2016) The post-embargo open access citation advantage: it exists (probably), it's modest (usually), and the rich get richer (of course). PLoS ONE 11(8):0159614
36. Amjad T, Sabir M, Shamim A, Amjad M, Daud A (2022) Investigating the citation advantage of author-pays charges model in computer science research: a case study of Elsevier and Springer. Libr Hi Tech 40(3):685–703
37. Langham-Putrow A, Bakker C, Riegelman A (2021) Is the open access citation advantage real? A systematic review of the citation of open access and subscription-based articles. PLoS ONE 16(6):0253129
38. Fraser N, Momeni F, Mayr P, Peters I (2020) The relationship between biorxiv preprints, citations and altmetrics. Quant Sci Stud 1(2):618–638
39. Momeni F, Dietze S, Mayr P, Biesenbender K, Peters I (2023) Which factors are associated with Open Access publishing? A Springer Nature case study. Quant Sci Stud 4(2):353–371
40. Hsu J-W, Huang D-W (2011) Correlation between impact and collaboration. Scientometrics 86(2):317–324
41. Puuska H-M, Muhonen R, Leino Y (2014) International and domestic co-publishing and their citation impact in different disciplines. Scientometrics 98(2):823–839
42. Sarigöl E, Pfitzner R, Scholtes I, Garas A, Schweitzer F (2014) Predicting scientific success based on coauthorship networks. EPJ Data Sci 3:1
43. Ni P, An X (2018) Relationship between international collaboration papers and their citations from an economic perspective. Scientometrics 116(2):863–877
44. Karimi F, Wagner C, Lemmerich F, Jadidi M, Strohmaier M (2016) Inferring gender from names on the web: a comparative evaluation of gender detection methods. In: Proceedings of the 25th international conference companion on World Wide Web, pp 53–54
45. Bornmann L, Mutz R (2014) From p100 to p100′: a new citation-rank approach. J Assoc Inf Sci Technol 65(9):1939–1943
46. Bornmann L, Williams R (2020) An evaluation of percentile measures of citation impact, and a proposal for making them better. Scientometrics 124(2):1457–1478
47. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd international conference on knowledge discovery and data mining, pp 785–794
48. Blasco BC, Moreno JJM, Pol AP, Abad AS (2013) Using the r-mape index as a resistant measure of forecast accuracy. Psicothema 25(4):500–506
49. Dong Y, Johnson RA, Chawla NV (2015) Will this paper increase your h-index? Scientific impact prediction. In: Proceedings of the eighth ACM international conference on web search and data mining, pp 149–158
50. Artur M (2021) Review the performance of the Bernoulli naïve Bayes classifier in intrusion detection systems using recursive feature elimination with cross-validated selection of the best number of features. Proc Comput Sci 190:564–570
51. Zhao L, Deng F, Zhang X, Yu N (2022) Rfe based feature selection improves performance of classifying multiple-causes deaths in colorectal cancer. In: 2022 7th International Conference on Intelligent Informatics and Biomedical Science (ICIIBMS), vol 7. IEEE, Los Alamitos, pp 188–194
52. Newbold P, Carlson WL, Thorne B (2013) Statistics for business and economics. Pearson Education, Upper Saddle River
53. Grech V, Rizk DE (2018) Increasing importance of research metrics: journal impact factor and h-index. Springer, Berlin
54. Egghe L (2006) Theory and practise of the g-index. Scientometrics 69(1):131–152
55. Bartneck C, Kokkelmans S (2011) Detecting h-index manipulation through self-citation analysis. Scientometrics 87(1):85–98
56. Asaro PM (2019) Ai ethics in predictive policing: from models of threat to an ethics of care. IEEE Technol Soc Mag 38(2):40–53
57. Zuiderveen Borgesius F et al (2018) Discrimination, artificial intelligence, and algorithmic decision-making. Línea. Council of Europe

Publisher's Note