EasyChair Preprint

№ 7502

Evaluation of Machine Learning to Early Detection of Highly Cited Papers

Galal Binmakhashen and Hamdi Al-Jamimi

February 25, 2022

Evaluation of Machine Learning to Early Detection of Highly Cited Papers

Galal M. BinMakhashen*, Hamdi A. Al-Jamimi**
Industry Collaboration Partnership VPRI
Interdisciplinary Research Center for Telecommunication Systems and Sensing
King Fahd University of Petroleum and Minerals
Dhahran, Kingdom of Suadi Arabia
*binmakhashen@kfupm.edu.sa , **aljamimi@kfupm.edu.sa

***Abstract*—As one of the fastest-growing topics, machine learning has many applications that span through different domains including image and signal recognition, text mining, information retrieval, robotics, etc. It enables information extraction and analysis for better insights and decision-based systems. The Web of Science(WoS) citation database is a leading organization that provides citation data of high-quality published research. WoS has its metrics to label published articles as Highly Cited Paper(HCP). Machine learning (ML) can help researchers in identifying the key characteristics of HCP. Moreover, it can allow research evaluation units forecasting significant scientific articles. In other words, it may allow researchers and/or research evaluators to detect potential scientific breakthrough ideas and stay current. In this study, more than 26 thousand records of published articles indexed by WoS were analyzed. All the records are drawn from the Technology research area as defined by WoS. Four ML algorithms are evaluated to verify the HCP common factors influence in raising citations and interest in scientific articles. The ensemble algorithms show promising results to identify HCP articles using only four factors.**

***Index Terms*—Highly-cited Research, Bibliometric Analysis, Machine Learning, Digital Libraries**

I. Introduction

There are several reasons behind a scientific article being cited by researchers and standout as highly cited. Researchers need to have their research reports appreciated by the research community and contributed to advance overall scientific knowledge.

The citation could be either positive or negative. The positive citations are happening when another research refer to a previously published in affirmative manner to advance his/her ongoing research. On the other hand, a formal expression of disagreement of the content of a previously published content can be considered as a negative citation. In this paper, we assume citations are all positive because negative citations are rare [1]. Consequently, the paper that receives high citation is representing a high quality paper, break-though or interesting topic to wide audience.

During the last two decades, an increasing interest was paid to the highly cited papers (HCP) analysis. It may allow research administration manipulate their funds strategies, and determine next research focus. Moreover, they were used as indicators for research assessments [2], [3]. In the context of scientific excellence in science policy, they were among the key indicators to identify and monitor scientific research outcomes. Recently, HCP is extended to measure scientific performance and research impact at institutions, universities, and countries levels [4].

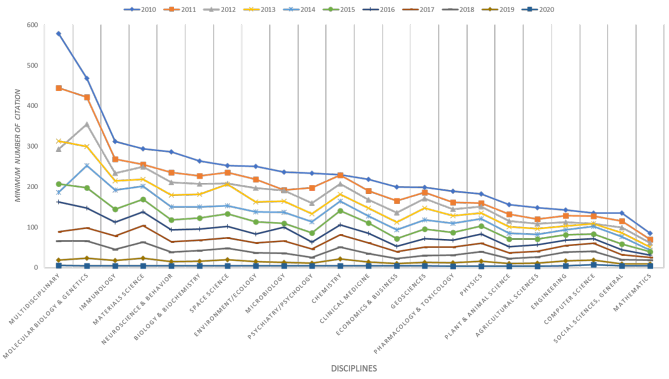

Fig. 1: ESI Highly Cited Paper Thresholds

On the contrary, it is not obvious what one is measuring using HCP. It made the application of mere citations count as an indicator a controversial with a level of uncertainty to measure scientific excellence [2]. Moreover, published papers with infrequent citations (i.e. lowly cited papers LCP) are not necessarily low quality than others and vice-versa [5]. There are other factors that affect such citations count of an article [6].

There are several studies that investigated the importance of the scientific articles' characteristics that affect their citation count. In general, they found that a paper with a research collaboration was more cited than others [2]. Moreover, the initial citation of an article is highly correlated with its long-term citations [7].

In this work, we grouped the features into two; pre-dissemination and post-dissemination features to study their affects on predicting which a paper will be highly cited. The study focuses on the automatic detection of HCP using WoS

The authors would like to acknowledge the help and support provided KFUPM University through funding the project number DF191012.

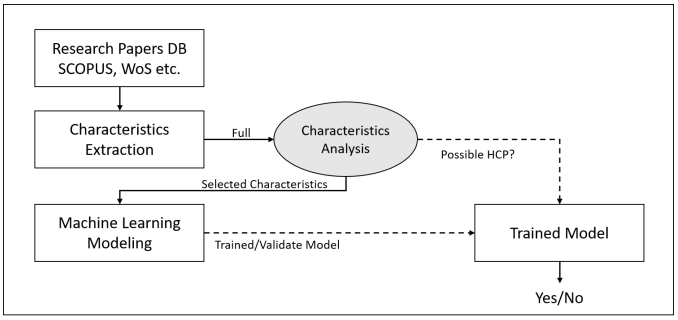

Fig. 2: High level HCP Detection Framework

databases as an example. The results can be extended later to other citation databases such as Scopus, Dimensions etc.

A. Highly Cited Criteria

There are two methods to what indicates highly cited articles: absolute and relative thresholds. These methods have been highlighted in different studies already and they were used to select the sample data analyzed in the previous studies [2], [8]–[10]. In this study, we collected the dataset from the Clarivate Analytics Essential Science Indicators (ESI). ESI labels a paper as a highly cited if it received enough citations to be placed in the top 1% of its research field per year. This makes the threshold more appealing as it is dynamic and related to a field of study. Figure 1 shows a research-field threshold in the year 2020 across several disciplines.

In this study, machine learning methods are investigated to build models that can forecast whether a recent article will receive enough citations to be the next highly-cited paper. Several features are extracted to build machine learning models.

B. Research Contributions

Motivated by the above observations, our study contributions are listed below:

- Analytically studying a set of important HCP characteristics.
- Constructing HCP detection framework using machine learning.
- Investigating several machine learning algorithms for identifying HPC papers using a set of articles' features while removing the current citation counts.

In Figure 2, the framework starts with collecting target papers metadata. Such data can be accessed through public electronic databases such as Clarivate (Web of Science). Once the target data is collected, a pre-processing is performed to extract characteristics per article metadata. The processed data, then, is analyzed statistically to identify significance to citation count. Moreover, it can be reduced by feature selection to retain significant characteristics. Finally, the processed data are modeled to build a good estimator for HCP detection.

This paper is organized as follows. Section II discusses the related work. Sections IV and presents the dataset, analysis, and the machine learning methods used in this study. Section IV-B discuss the study findings and results. Finally, Section V presents our conclusions.

II. Related work

A. Characteristics of highly cited papers

Usually, such characteristics are grouped into three levels; paper, journal, and author characteristics.

Several studies have shown a significant correlation between published paper characteristics and its received citations count [11]. Among these characteristics the title length, abstract length, article's pages. Moreover number of keywords, references, tables, and figures are used for analysis. All these features are representing a pre-dissemination characteristics where authors may consider to improve their articles' chances to be cited.

Another important factor is the journal impact factor (IF) or CiteScore(CS). Such factor may be viewed as a measure, not only for the journals' quality, but the papers' quality as well. Hence, the higher the IF or CS of a journal the higher the quality of the paper. Consequently, IF and/or CS are contributing articles' citations count. There are some studies that have provided a strong statistical evidence that the publishing in journals with high IF/CS are positively correlated to citation count [12], [13]. Moreover, the journal's coverage and scope are among the significant factors for HCP. For instance, articles published in multi-disciplinary journals are expected to receive more citations than other published papers in specialized journals [13]. In Figure 1, it shows that a multi-disciplinary article requires more than 500 citations to be HCP. Similarly, journals' coverage has either a positive or negative role to promote the published papers relatively to the journals' audience. It is obvious that national journals receive fewer citations than their equivalent international journal [14].

International research is not the property of journals' coverage only. Usually, authors play a key role to lead such research. Authorship collaboration can be used to indicate the extent of national or international research [12]. The international collaboration in research has been identified as a good factor for published papers to receive more attention [12]. For instance, Aksnes et. al [2] observed that the number of authors contributing to the HCP is larger than ordinary published papers.

In general, the number of authors feature demonstrated a considerable impact on the paper's citation. Noorhidawati et al [15] reported most of HCPs had two to five authors and the fraction of papers authored by ten researchers or more was about 25% in the study. Additional studies stated that there is a positive correlation between the number of authors and the paper's impact (i.e. citation count) [11], [16].

There are several research questions that have been investigated previously; such as how high impact is using citation counts in case of academia-corporate research collaboration [17]. Does the impact of authors count on article citation equally likely across research-fields [16]. Moreover, there are other author-based factors that have been investigated previsouly such as the authors' academic rank, productivity,

TABLE I: Summary Related Works

Study	Obj	Method	F(#)	Domain	Data
[10]	BI	RM	M	Biology & genetic	WOS(1995-2005)
[28]	HCP	PCA	6	Multidisciplinary	MDPI(2017)
[25]	LTCI	RNN	16	Multidisciplinary	WoS(1980-2003)
[29]	RA	SA	4	Computer Science	DBLP
[21]	HCP	NN	4	physics	APS-ESI (1980-1989)
[30]	CCP	RM	16	physics	KDD CUP (1993 - 2003)
[31]	LTCI	RM	2	physics	WoS
[7]	BI	SA	1	physics	WOS (1995 - 2012)
[32]	HCP	SA	1	physics & biology	NA
[27]	HCP	SA,NN	4	biomedical	PubMed
[23]	HCP	Fuzzy	25	Multidisciplinary	SCI
[26]	CCP	RM, NN	3	Computer Science	ArnetMiner

Obj: Objectives, BI: Breakthrough, Identification,Cite: citations, chars: characters, CCP: Citation Count Prediction, Dyn: Dynamics, ESI: Essential Science Indicators,F: Factors, LTCI: Long-Term Citation Impact, Lang: Language. RM: Regression Model, PCA: Principal Component Analysis, Recurrent Neural network, NA: Not Applicable, NN: Neural Networks, SA: Statistical Analysis, Fuzzy: Fuzzy Logic methods.
APS: American Physical Society, M: Many Factors

reputation, etc. as in [16], [18].
Often, there are different citation thresholds for each research field as computed by WoS. Noorhidawati et al. [15] reported that more than 50% of the HCPs belong to technology and engineering field, and only 16% represented medicine. Another study stated that the citations received by papers on social science are higher than those published on natural science [18]. Moreover, the citations may be affected by field size. For instance, papers published on organic chemistry, analytical chemistry, and physical chemistry received higher citations than those published on biochemistry [19]. It is expected that the hot topics would attract more attention and receive more citations accordingly [20].

III. Machine Learning & HCP

In general, there are two paradigms for forecasting HCP in the literature. First, researchers used a full set of features that spans paper, journal, author features to forecast citations. The features are either manually or analytically computed from samples of highly cited or Nobel prize research data [5]. Then, a regression method is used for forecasting the future citation count of a focal paper [10]. Secondly, researchers used feature selection methods and highlight a subset of the HCP characteristics that helps machine learning algorithms to reduce training time [21].
By setting up high citations as a response variable, scientists have tried to predict important phenomena such as breakthrough research, new areas of research, long-term scientific impact, etc. [21]. Table I summarizes studies that used machine learning to address various research problems.
Ponomarev et al. [10] developed forecasting models to identify breakthrough candidate publications by predicting the future citation patterns using time dependent analysis of citation rates. The top ranking in citations is a good proxy for measuring the impact of research; however, it is not a sufficient condition to consider the paper as breakthrough research. Thus multidimensional feature space could be used as in [22]. Another work by Wang et al. in [23] proposed a prediction model that used 25 features to forecast citations into low, medium, or high. The developed model was giving predictions within 15 years. It indicated that the first author, paper's quality, and reputation of the journal were the most relevant predictors for high citation. Moreover, Wang et al. [24] integrated both bibliometric and altimetric factors for predicting the publication citation growth. Among the identified important factors was the influence of the first author. Another study considered the early citation of the paper for long-term citations [25]. However, long-term citation might not be sustained if breakthrough research is identified. Therefore, an early citation may not be effectively indicates the long-term citation. Still, the factors that are influencing the paper citation growth or sustaining its long-term citation is not thoroughly identified. Another study [26] identified that the author expertise and venue have the strongest impact on the citation predictions.
In contrast to the above findings, Hurley et al. [27] reported journal and language characteristics are more important than the number of authors/co-authors that influence citation behavior. They derived their conclusion using logistic regression models.
In summary, identifying a set of effective features may help researchers to identify a list of recommendations to shape their scientific reports accordingly and draw attention towards their findings.
In this study, we are evaluating a machine learning methods for predicting highly cited papers using 16 features. Unlike the above studies where the task is regression to predict the long-term citations count. In this study, we focus more on a set of article's features that can be utilized for HCP. As we assume that the HPC features can be used by researchers to shape their scientific publications before dissemination. Moreover, the most effective articles features are highlighted for HCP. The data has two labels either highly cited or lowly cited. The labeling of the data was done according to ESI where HCP papers are marked with a value of 1.

A. Background Machine Learning Methods

In this section, we are presenting a brief background of the machine learning algorithms adopted in this study.

- *Support Vector Machines:*
  Support Vector Machines (SVM) is a discriminative classifiers that assumes a clear separation between any two groups of data [33]. So, the key task of SVM is to maximize the separation between these classes. However, a soft margin SVM is used to relax the clear-separation assumption by SVM and allow classes overlapping. This is done by optimizing the following equation:

$$\begin{aligned} &\min_{w,\xi} 0.5 w^T w + C \sum_{i=1}^{n} \xi_i \\ &s.t. w^T x_n t_n \geq 1 - \xi_n \quad \forall_n \\ &\xi_n \geq 0 \quad \forall_n \end{aligned} \tag{1}$$

  where $t \in \{-1, 1\}$ and ξ_n are penalties for those points violates the decision margin.
  Finally, after training and finding the best parameters, the classification outcome is computed by:

$$\arg\max_{t} (w^T X_{test}) t \tag{2}$$

For non-separable data, a kernel trick was introduced by Cortes and Vapnik in [34] to transform the data to another dimension space that could be separable. Therefore, Equation 3 is modified as follows:

$$\arg\max_{t}(w^T\phi(X_{test}))t \tag{3}$$

where $\phi(.)$ is called a kernel function.

- *K-Nearest Neighbor:*
  K-Nearest Neighbor (kNN) is an instance-based classifier which determines the query label based on evidence of the closest samples in the training set. Moreover, it is a non-parametric algorithm which means no initial strong assumptions about the classification space should be made before evaluation [35]. Formally, a test point: $\mathbf{x}$ define a set of the k nearest neighbors of $\mathbf{x}$ as $S_{\mathbf{x}}$, where $S_{\mathbf{x}} \subseteq D$ s.t. $|S_{\mathbf{x}}| = k$ and $\forall(\mathbf{x}', y') \in S_{\mathbf{x}}$,

$$dist(x, x') \leq \min_{(x'',y'')\in S_x} dist(x, x'') \tag{4}$$

  A classifier $kNN()$ is defined as a function returning the common label of samples in $S_{\mathbf{x}}$:

$$kNN(x) = mode(y'' : (x'', y'') \in S_x), \tag{5}$$

  where $mode(\cdot)$ is returning the most frequent label in the $S_{\mathbf{x}}$.
- *Trees Classifier:*
  - Decision tree method
    A decision tree method classifies (i.e. categorizes) a data instance by finding the fittest rule down the tree (i.e. from root to leaf nodes) that allows such an algorithm to produce a decision. The algorithm refines its decision for the given features of an instance by repeatedly selecting a branch/sub-decision at each point (i.e. node). A final decision (class label) is produced once the algorithm reaches the leaf nodes. A mid-node may have at least two branches (children nodes) where the leaf node does not have any children. The Decision tree algorithm is straightforward, but in terms of its structure, the number of nodes can be gigantic.
    In this work, a decision tree is built using the Gini impurity method. Suppose we have C classes and $p(i)$ is the probability of picking an instance with a class label i, then the Gini Impurity is calculated as:

$$G = \sum_{i=1}^{C} p(i) \times (1 - p(i)) \tag{6}$$

  - Random Forest
    The Random forest is among well-known machine learning algorithms [36]. The algorithm constructs multiple uncorrelated decision trees uses bootstrap aggregation (bagging) technique [37]. So, Random Forest, as its name implies, it consists of a large number of individual decision trees that operate as an ensemble classifier. Each individual tree in the random forest constructors refines rules for class prediction. Then, the most voted label by all trees becomes the class label.

TABLE II: Features Definitions

Category	Feature	Definition
2*Author-based	AU	Authors count
	COL	Inter. collaboration(True, False)
Project-based	FUN	Research funding(True, False)
8*Article-based	DT	Title length
	TP	Title punctuation count
	DTY	Document type
	KW	Keywords count
	ABS	Abstract length of characters
	REF	References count
	SP	Supplement (True, False)
	PC	Page count
5*Journal-based	PUB	Publisher
	SI	Special issue (True, False)
	LNG	Language
	IF	Journal Impact factor
	QR	Journal quartile

- *Adaboost Classifier:*
  The AdaBoost classifier is an ensemble learning method. Unlike other ensemble methods, AdaBoost starts with a classifier to model the whole training data. Then, it uses multiple instances of the same classifier to model the incorrectly classified instances of the first classifier. The method adopted in this work is called Adaboost SAMME and described in this work [38].
- *Naive Bayes Classifier:*
  Naïve Bayes classifier is based on Bayes theorem with an assumption that all given features are independent. So, the Naïve Bayes classifier assigns the most likely class label to a given instance by the following:

$$P(y|X) = \frac{P(X|C)P(C)}{P(X)}, \tag{7}$$

  where $X = (x_1, x_2, ..., x_n)$ is an instance of a feature vector, C is a class label, $P(y|X)$ is a posterior probability, $P(X)$ is a proir probability of predictor, and $P(C)$ is a proir probability of class. The naïve Bayes classifier is very successful in many domains [39].

IV. Experimental Work

A. Data collection

The database used in this paper covers highly cited papers (Technology research area) as defined by the Clarivate ESI during the period from 2009 to 2019. The data collected from ESI contains metadata about 26154 highly cited papers, 252,015 authors, and 7,036,905 citations. ESI is a prominent platform that deals with consistent evaluation indicators for universities, governments, and research institutions. It offers unbiased metrics that measures 22 academic fields and capturing six indices; the total number of cited papers, citation frequency, the mean value of citation per paper, highly cited papers, hot papers, and also top papers.
To have a balanced dataset we also collected metadata about the similar number of lowly cited papers (according to ESI definition) in the same period 2009-2019. These papers are

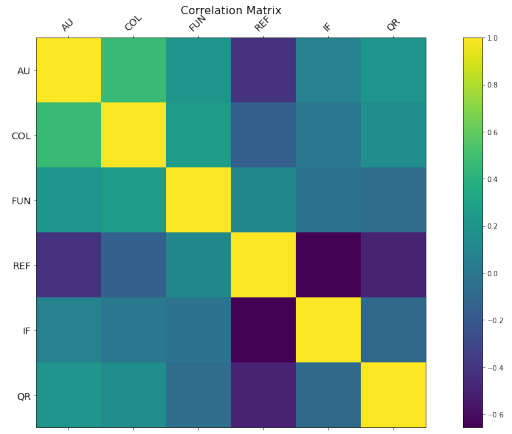

Fig. 3: Six Features correlation matrix

authored by 170,073 researchers and received 657,339 citations.

B. Results and Discussion

1) Identify Highly Cited Paper Features: Present studies show that paper citations are influenced by several features that can be categorized by relevance to author, field, article or journal features [5], [12], [19]. In this study, an initial set of 16 features is extracted from the ESI metadata to represent the articles. The features belong to four general categories as listed in Table II. Spearman correlation analysis is used to investigate the relations between the features and scientific impact. Six features are found significant as depicted on Figure 3.

2) ML methods settings: We compared four machine learning algorithms; two ensemble methods (RandomForest, AdaBoost), a discriminative method (SVM), and a generative method (Naive Bayes). The algorithms are configured and tuned to predict HCP.

The RandomForest method is configured with a maximum of 100 trees. The trees were generated randomly to fit the training data using the Gini Impurity method. The combination of these trees is used to make a final decision. For AdaBoost, we adopted the SAMME.R implementation for faster convergence. A radial basis function (RBF) is used for SVM to model the training data. The SVM, then, is configured with C = 1.0 and gamma =1.0. The setting of SVM was computed empirically.

3) Results and Performance Evaluation: The results are assessed using the standard machine learning metrics; Precision, Recall, and f1-score. Equations (8,9, and 10) formulate these metrics.

$$Precision = \frac{TP}{(TP + FP)} \tag{8}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{9}$$

$$f1score = 2 \times (\frac{Precision \times Recall)}{(Precision + Recall)} \tag{10}$$

TABLE III: Experiments Results using (16) Features

Methods	***Precision***	***Recall***	***F1-Score***
Random Forest	0.89	0.89	0.89
AdaBoost	0.82	0.82	0.82
SVM	0.74	0.53	0.4
Naive Bayes	0.73	0.66	0.63

TABLE IV: Experiments Results using Four Features only

Methods	***Precision***	***Recall***	***F1-Score***
Random Forest	0.82	0.82	0.82
AdaBoost	0.8	0.8	0.8
SVM	0.78	0.78	0.78
Naive Bayes	0.72	0.7	0.7

where TP, FP, and FN are the true positive, false positive and false negative of the calculated results.

Tables (III, IV) summarize the experiments results.

The results illustrated in Tables (III, and IV) are showing that the ensemble methods outperform the others. Using RandomForest, the model was able to generalize and predict a paper to be highly-cited correctly with 89% $f1score$ using 16 features. Moreover, the precision and recall were the same. SVM and Naive Bayes classifiers were performed the worst due to the high overlap among the classes. As the metadata collected for lowly-cited articles in several examples are found similar to those of the highly-cited, these two classifiers made many miss-classifications (Table III).

As we can observe from the matrix in Figure 3, there are two features that are highly correlated. Therefore, we removed them from the feature space and rebuild ML models. The results of the RandomForest classifier is negatively affected by this selection of features and the $f1score$ reduced from 89% to 82%. Similarly, the performance of the AdaBoost method was reduced from 82% to 80% $f1score$ in both experiments respectively.

On the other hand, the performance of both the SVM and Naive Bayes classifiers were improved by reducing the features from 16 to 4. The SVM model was able to improve from 40% to 78% $f1score$, and Naive Bayes was reached

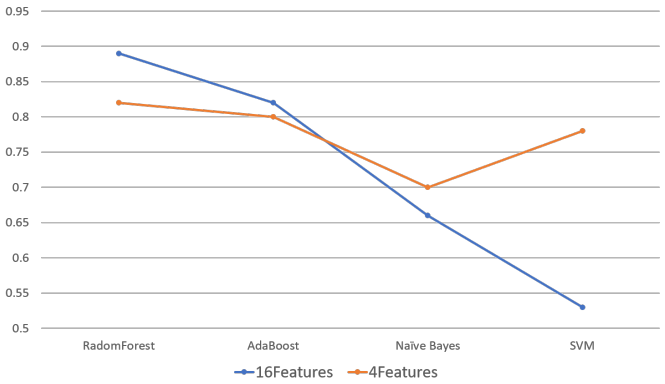

Fig. 4: Machine Learning Method Accuracy

70% $f1score$.

One reason for the above performance is that these two algorithms assume feature independence. Therefore, removing correlated features has helped in improving the performance of these two ML algorithms. To sum up, based on the experiments conducted in this study the number of authors(AU), Research Funding info(FUN), international collaboration(COL), and journal quartile(QR) are found the most important characteristics for HCP.

V. Conclusion

Usually, highly cited papers are sharing common characteristics. Such characteristics are very important to consider raising the visibility of the scientific communication of researchers. On the other hand, some of these characteristics are dynamic and changing from time to another. In this work, we investigated the highly-cited papers' factors according to the information extracted from Clarivate Web of Science database. Among all factors, we selected the common 16 features that match previous literature. The statistical analysis reveals that 6 of the factors are playing an important role in the class of the paper (whether it is highly or lowly cited). However, some of these factors are inter-correlated. Furthermore, the experiments showed that researchers' may consider only four features as important. In the future, we will consider applying a similar methodology using another common research database such as Elsevier-SCOPUS and compare the results.

References

[1] C. Catalini, N. Lacetera, A. Oettl, The incidence and role of negative citations in science, Proceedings of the National Academy of Sciences 112 (45) (2015) 13823–13826.

[2] D. W. Aksnes, Characteristics of highly cited papers, Research evaluation 12 (3) (2003) 159–170.

[3] A. F. Van Raan, The pandora's box of citation analysis: measuring scientific excellence, the last evil, The web of knowledge: A festschrift in honor of Eugene Garfield (2000) 301–319.

[4] A. Zeng, Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang, H. E. Stanley, The science of science: From the perspective of complex systems, Physics Reports 714 (2017) 1–73.

[5] I. Tahamtan, A. S. Afshar, K. Ahamdzadeh, Factors affecting number of citations: a comprehensive review of the literature, Scientometrics 107 (3) (2016) 1195–1225.

[6] L. Bornmann, R. Williams, et al., An evaluation of percentile measures of citation impact, and a proposal for making them better, Scientometrics (2020) 1–22.

[7] J. Winnink, R. J. Tijssen, Early stage identification of breakthroughs at the interface of science and technology: lessons drawn from a landmark publication, Scientometrics 102 (1) (2015) 113–134.

[8] O. Persson, Are highly cited papers more international?, Scientometrics 83 (2) (2010) 397–401.

[9] J. Antonakis, N. Bastardoz, Y. Liu, C. A. Schriesheim, What makes articles highly cited?, The Leadership Quarterly 25 (1) (2014) 152–179.

[10] I. V. Ponomarev, D. E. Williams, C. J. Hackett, J. D. Schnell, L. L. Haak, Predicting highly cited papers: A method for early detection of candidate breakthroughs, Technological Forecasting and Social Change 81 (2014) 49–55.

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

[12] F. Didegah, M. Thelwall, Which factors help authors produce the highest impact research? collaboration, journal and document properties, Journal of informetrics 7 (4) (2013) 861–873.

[13] J. K. Vanclay, Factors affecting citation rates in environmental science, Journal of Informetrics 7 (2) (2013) 265–271.

[14] B. Millet-Reyes, The impact of citations in international finance, Global Finance Journal 24 (2) (2013) 129–139.

[15] A. Noorhidawati, M. Y. I. Aspura, M. Zahila, A. Abrizah, Characteristics of malaysian highly cited papers, Malaysian Journal of Library & Information Science 22 (2) (2017) 85–99.

[16] C. Biscaro, C. Giupponi, Co-authorship and bibliographic coupling network effects on citations, PloS one 9 (6) (2014) e99502.

[17] C. Bloch, T. K. Ryan, J. P. Andersen, Public-private collaboration and scientific impact: An analysis based on danish publication data for 1995–2013, Journal of Informetrics 13 (2) (2019) 593–604.

[18] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, A. Mukherjee, Towards a stratified learning approach to predict future citation counts, in: IEEE/ACM Joint Conference on Digital Libraries, IEEE, 2014, pp. 351–360.

[19] L. Bornmann, H. Schier, W. Marx, H.-D. Daniel, What factors determine citation counts of publications in chemistry besides their quality?, Journal of Informetrics 6 (1) (2012) 11–18.

[20] M. J. Gallivan, Analyzing citation impact of is research by women and men: do women have higher levels of research impact?, in: Proceedings of the 50th annual conference on Computers and People Research, 2012, pp. 175–184.

[21] F. Wang, Y. Fan, A. Zeng, Z. Di, Can we predict esi highly cited publications?, Scientometrics 118 (1) (2019) 109–125.

[22] I. V. Ponomarev, B. K. Lawton, D. E. Williams, J. D. Schnell, Breakthrough paper indicator 2.0: can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction?, Scientometrics 100 (3) (2014) 755–765.

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

[24] M. Wang, Z. Wang, G. Chen, Which can better predict the future success of articles? bibliometric indices or alternative metrics, Scientometrics 119 (3) (2019) 1575–1595.

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

[26] R. Yan, J. Tang, X. Liu, D. Shan, X. Li, Citation count prediction: learning to estimate future citations for literature, in: Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, pp. 1247–1252.

[27] L. A. Hurley, A. L. Ogier, V. I. Torvik, Deconstructing the collaborative impact: Article and author characteristics that influence citation count, Proceedings of the American Society for Information Science and Technology 50 (1) (2013) 1–10.

[28] M. Elgendi, Characteristics of a highly cited article: A machine learning perspective, IEEE Access 7 (2019) 87977–87986.

[29] N. Malik, H. U. Khan, M. S. Faisal, A. Mahmood, S. Seo, M. R. Bhutta, A novel approach for finding research areas for new researchers, New Review of Hypermedia and Multimedia 25 (3) (2019) 182–204.

[30] J. Chen, C. Zhang, Predicting citation counts of papers, in: 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), IEEE, 2015, pp. 434–440.

[31] C. Stegehuis, N. Litvak, L. Waltman, Predicting the long-term citation impact of recent publications, Journal of informetrics 9 (3) (2015) 642–657.

[32] M. Newman, Prediction of highly cited papers, EPL (Europhysics Letters) 105 (2) (2014) 28002.

[33] C.-W. Hsu, C.-J. Lin, A comparison of methods for multiclass support vector machines, IEEE transactions on Neural Networks 13 (2) (2002) 415–425.

[34] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[35] N. S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, The American Statistician 46 (3) (1992) 175–185.

[36] Z. Lv, S. Jin, H. Ding, Q. Zou, A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features, Frontiers in bioengineering and biotechnology 7 (2019) 215.

[37] A. Liaw, M. Wiener, et al., Classification and regression by randomforest, R news 2 (3) (2002) 18–22.

[38] T. Hastie, S. Rosset, J. Zhu, H. Zou, Multi-class adaboost, Statistics and its Interface 2 (3) (2009) 349–360.

[39] I. Rish, et al., An empirical study of the naive bayes classifier, in: IJCAI 2001 workshop on empirical methods in artificial intelligence, Vol. 3, 2001, pp. 41–46.