

Forecasting Publication's Success Using Machine Learning

Citation for published version (APA):

Alchokr, R., Haider, R., Shakeel, Y., Leich, T., Saake, G., & Krüger, J. (2023). Forecasting Publication's Success Using Machine Learning. In I. Frommholz, P. Mayr, G. Cabanac, S. Verberne, & J. Brennan (Eds.), *BIR 2023 : Bibliometric-enhanced Information Retrieval: Proceedings of the 13th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 45th European Conference on Information Retrieval (ECIR 2023)* (pp. 77-89). (CEUR Workshop Proceedings; Vol. 3617). CEUR-WS.org.
<https://jacobkrueger.github.io/assets/papers/Alchokr2023ForecastingSuccess.pdf>

Document license:

CC BY

Document status and date:

Published: 01/01/2023

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Forecasting Publications' Success Using Machine Learning Prediction Models

Rand Alchokr^{1,*}, Rayed Haider^{3,*}, Yusra Shakeel^{1,2,*}, Thomas Leich^{3,4}, Gunter Saake¹ and Jacob Krüger⁵

¹Otto-von-Guericke University, Magdeburg, Germany

²Karlsruhe Institute of Technology, Karlsruhe, Germany

³Hochschule Harz, Wernigerode, Germany

⁴METOP GmbH, Magdeburg, Germany

⁵Eindhoven University of Technology, The Netherlands

Abstract

Measuring the success and impact of a scientific publication is an important, thus controversial matter. Despite all the criticism, it is widespread that citation counts is considered a popular indication of a publication's success. Therefore, in this paper, we use a machine learning framework to test the ability of alternative metrics (altmetrics) to predict the future impact of papers reflected in the citation counts. To achieve the experiment, we extracted 7,588 papers from 10 computer science journals. To build the feature space for the prediction problem, 14 different altmetric indices were collected, 3 feature selection approaches, namely, Variance threshold, Pearson's Correlation, and Mutual information method, were used to minimize the feature space and rank the features according to their contribution to the original dataset. To identify the classification performance of these features, three classifiers were used: Decision Tree, Random Forest, and Support Vector Machines. According to the experimental data, altmetrics can predict future citations and the most useful altmetrics indications are social media count, tweets, news count, capture count, and full-text view, with Random Forest outperforming the other classifiers.

Keywords

Bibliometric, alternative metrics, machine learning, computer science

1. Introduction

A successful publication is a desirable goal for any researcher, irrespective of their scientific field. However, judging how successful a published paper is and measuring that success is considered a critical issue. Furthermore, forecasting scientific impact and success is becoming an essential regular task for the hiring committees, funding agencies, and department heads for recruitment decisions and rewards [3, 7, 26]. Through this a merit-based career advancement scheme is developed, that forecasts the individual's performance based on past achievements and projects future performance. However, distilling the contents of each article into an appraisal

BIR 2023: 13th International Workshop on Bibliometric-enhanced Information Retrieval at ECIR 2023, April 2, 2023

*Corresponding author.

✉ rand.alchokr@ovgu.de (R. Alchokr); rayedhaider95@gmail.com (R. Haider); yusra.shakeel@kit.edu (Y. Shakeel); tleich@hs-harz.de (T. Leich); saake@ovgu.de (G. Saake); j.kruger@tue.nl (J. Krüger)

🆔 0000-0003-0112-5430 (R. Alchokr); 0000-0001-5135-4325 (Y. Shakeel); 0000-0001-9580-7728 (T. Leich); 0000-0001-9576-8474 (G. Saake); 0000-0002-0283-248X (J. Krüger)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

of an individual's past, present, and future influence and determining an acceptable ranking of candidates is significantly challenging when presented with candidate pools ranging from a few hundred for tenure-track positions to thousands for fellowship and grant competitions.

In the past decades, researchers have relied heavily on quantitative indicators for evaluating the scientific success of a given research body. Citation frequency is a well-known criterion for research evaluation and despite all the criticisms of citations not being a perfect and objective means of measuring scientific quality, citation counts are still widely referred to as a foremost indicator of the impact and success of a publication in the scientific community. Recently, there has been extensive research to investigate the link between a paper's citations and all possible factors correlated to it [5, 10, 25, 23, 24, 15, 6]. Additionally, research is becoming interested in forecasting the future success of a paper [7, 13, 38, 39, 2]. Among these factors, bibliometrics and altmetrics are deemed of the utmost relevance. While bibliometrics is the traditional indices reflecting the characteristics as well as the credibility of papers, authors, and publishing venues (e.g. citations, h-index of the author, cite score of the venue, etc.), altmetrics have been recently introduced to capture the spread of a publication in various online platforms (e.g. Wikipedia, Twitter, Facebook). The combination of both bibliometrics and altmetrics, is recommended by researchers to complement their pros and cons [23]. Recent studies have looked at the relationship between bibliometric indicators and altmetrics, taking into consideration peer-reviewed quality evaluation methods [24, 8, 27, 33, 32, 30, 31]. The application of altmetrics in research assessment raises the question of whether the data collected by altmetrics is a good predictor of future success and whether it correlates with citation.

On the other hand, the remarkable progress in the field of machine learning (ML) has matured a plethora of techniques that can efficiently handle various forecasting tasks. In the context of predicting papers' citations using bibliometrics and altmetrics, multiple studies formulate the problem into a regression task that considers continuous values of both features and the output [13, 2, 17, 22, 29], whereas other studies considered classification algorithms that generate categorical outcome [38, 39]. In this paper, we rely on both of these metrics to find out which altmetrics features contribute to forecasting citation counts. *We consider a paper successful if it achieves a high number of citations.* We categorize the publications according to their citation counts. Belonging to a class of higher ranking hints at a more successful paper. The goal of this study is to determine which features of altmetric are useful in predicting future highly cited papers and which machine learning model would be the best for this prediction. In our experiment, we will use Decision Trees, Random Forests, and Support Vector Machines.

In detail, our main contributions in this paper are as follows:

- We collect an extensive dataset comprising papers from 10 computer engineering journals from 2010 to 2015. Further, we elicit the papers' citations and altmetrics, aiming to find the most promising altmetrics formula to predict the future success of a paper.
- We discuss multiple prediction models and compare their accuracy.

Through our experiments, we aim to provide a better understanding of the usefulness of altmetrics to indicate the future success of publications.

2. Background

Next, we present the background needed to understand this paper.

2.1. Evaluation Metrics

Peer Reviewing during the scientific evaluation process of papers is an essential part of publishing academic research, representing an important quality assurance mechanism [34]. On the other hand, bibliometrics which represent the traditional metrics are common measures that the research community relies on when assessing the scientific impact and quality of a publication [10]. Such metrics have multiple advantages, they mainly facilitate the examination of large datasets and help decision-making on individuals, institutions, or research grants [21]. Citation counts, h-index, and impact factor are among the most important metrics used for assessing the impact and quality of publications, publishing venues, authors, or research in general. Citation-based metrics are assumed to directly reflect on the impact and quality of a publication by implying credibility to the reader and reflecting the total impact of a publication on a research field [25]. Despite their potential benefits, bibliometrics have always been criticized in the context of measuring the impact or quality of research, which they do not necessarily capture [21]. However, many studies suggest that using bibliometrics is a helpful complement to mitigate potential biases during traditional peer review.

Altmetrics have been recently introduced as means to assess the impact of a publication based on publicly available interfaces of various online platforms [18]. These metrics allow researchers to track the impact of publications beyond traditional bibliographic metrics and help them in catching the buzz and spread of their research to a broader audience by calculating quantitative values of user interactions on social platforms, for instance, Wikipedia, Twitter, Facebook, the number of downloads, views, or read times. It is known that altmetrics may not accurately represent scientific quality, they lack the evidence, are difficult to measure, commercialized, and easily manipulated [36, 23], however, based on the mentioned benefits, many researchers argue that altmetrics can serve as an impact indicator and a complement to traditional metrics [23, 24, 15]. Researchers recommend using both kinds of metrics when assessing the impact or quality of a publication to complement their pros and cons [23].

In conclusion, we rely on both kinds of metrics to measure the success of a publication. We consider a paper successful and has an impact if it has achieved a high number of citations.

2.2. Predictive Algorithms

By definition, machine learning is a branch of computer science that grew out of artificial intelligence research into pattern recognition and computational learning theory [14]. It is the learning and building of algorithms that can learn from and make predictions on datasets. There are three types of machine learning algorithms: 1) Supervised learning algorithms: with two types: classification and regression, 2) Unsupervised learning algorithms: association, clustering, and dimensionality reduction, and 3) reinforcement learning.

Supervised learning, is defined as learning from labeled training data. The training data is learned using a supervised learning algorithm, which then creates a prediction function. For

unseen occurrences, the predictive function will be utilized to determine the class label. Linear regression, Logistic Regression, CART, Naïve Bayes, and K-Nearest Neighbors (KNN) — are examples of supervised learning. Also, Bagging with Random Forests, Boosting with XGBoost, and Multilayer Perception (basic ANN). To start with, Naive Bayes applies the assumption of independence between every set of features, meaning that all features contribute independently to the probability of the target's outcome [16]. XGBoost is a scalable tree-boosting system, it is used widely by data scientists nowadays [11]. A classifier is an example of a supervised learning algorithm. Machine learning algorithms that tackle the categorization problem are known as classifiers.

A classification issue is described as a task of determining class labels for fresh observations based on a training batch of data with a known class label. ANN is a helpful model for classification, clustering, pattern recognition, and prediction in many fields [1]. Random Forest inputs and random features produce good results in classification—less so in regression. Finally, The K-Nearest Neighbors (KNN) has often been used in pattern recognition problems. According to the existing literature, there have been various studies done to investigate the factors that influence citation and studies that attempted to forecast and estimate future citations.

3. Related Work

According to existing literature, various studies investigated the factors that influence citation, while others attempted to forecast and estimate future citations. Some of these studies utilized the early citation counts to predict the publication's future success [35, 2, 29]. Their results agree on the impact early citations and other related factors have on predicting highly cited publications. Social media metrics started to gain interest in research. For instance, tweets had a weak ability to positively predict high citation counts across several disciplines [20]. In the computer science domain, multiple classification methods were used to check whether the future success of articles depends on bibliometrics or altmetrics, and the results show that both contribute equally with PCA achieving the best performance [39]. Another study investigated altmetrics specifically using the "Altmetric Attention Scores", but this time to predict the retraction of the articles. The results show that roughly one-fourth of the retractions are properly predicted using five alternative metrics Copiello [12]. Another study by Akella et al. [4] used atmetrics social media features to predict early and long-term citation counts using several classifiers and regressors, their main results indicate that Mendeleev readership plays a crucial role in determining the early citations. We built our experiments on theirs, but first by determining the most influential features.

We present an overview of the related work in Table 1 collected by conducting a literature search on Scopus¹ digital library. For each study, we display the type of prediction, and feature selection methods that provide the necessary background information to guide our experiments. Overall, the literature demonstrates that researchers have explored a variety of machine learning algorithms and features in their efforts to predict the academic influence of research publications.

¹<https://scopus.com>

Table 1
Overview of the related work.

Ref	Algorithms	Results	Fields
Wang et al. [39]	Classification (Naïve Bayes, KNN, Random Forest), Relief-F, Principal Component Analysis (PCA) and entropy weighted method to find which better predict the future success of articles: Bibliometrics or altmetrics	PCA has the best performance with 0.947 precision	CS
Poggi et al. [27]	Classification Correlation-based Feature Selection (CFS)	SVM outperforms the other classification methods with 0.894 precision	CS
Bornmann et al. [8]	Regression calculated an adjusted R2, journal papers published in 1980 Journal impact, number of authors, the number of cited references, and the number of pages	The consideration of journal impact improves the prediction of long-term citation impact	CS,WoS
Copiello [12]	Classification comparing a set of 100 retracted articles with high Altmetric Attention Scores with a sample of 100 randomly chosen retracted by PLoS ONE	Roughly one-fourth of the retractions are properly predicted using five alternative metrics	CS
Akella et al. [4]	Classification, Multiple Linear Regression, altmetrics social media features to predict the early and long-term citation counts	neural networks and ensemble models performed better, with high predicted accuracy and F-1 scores. Mendeley readership plays a crucial role in determining the early citations	CS
Bai et al. [7]	Paper Potential Index (PPI) model and multi-feature model	PPI model outperforms the multi-feature model in terms of range-normalized RMSE and it better interprets changes in citation without requiring parameter adjustments. In terms of Mean Absolute Percentage Error and Accuracy, the multi-feature model outperforms the PPI model; nevertheless, its predictive performance is more dependent on parameter modification	CS,MS
Yu et al. [40]	Stepwise multiple regression used to select appropriate features and to build a regression model for explaining the relationship between citation impact and the chosen features (external features of a paper, authors, journal, citations)	Regression model works well in this situation where bibliometrics have high predictability compared to other features and that the regression model works well in this situation	InfS,LibS
Hassan et al. [20]	Linear regression, sentiment analysis of the publications tweets (positive, negative, neutral) of 6,482,260 tweets, July 2011 to June 2016, user's profile, types of journals, citation count, subjects	A weak positive prediction of high citation counts across 16 broad disciplines in Scopus, number of unique Twitter users improved the adjusted R-squared value of regression analysis in several disciplines	CS
Stegehuis et al. [35]	Quantile regression, utilized citations to predict publication's future success (Impact factor of the publication and the First 1-year citation counts) are used as predictors	Both predictors (i.e., impact factor and early citations) contribute to the accurate prediction of long-term citation impact	P
Daud et al. [13]	CART, Naive Bayes, Maximum Entropy Markov, bibliometrics: author, co-author, venue of publication	Maximum Entropy Markov model had a better prediction of the average number of citations whereas CART performed better for predicting an average relative increase in citations. They concluded that an excellent paper will be cited regardless of the paper's publishing time and a high-quality paper will have a high influence	CS
Fu and Aliferis [17]	Logistic regression, support vector machine modules, cross-validation, AUC, HITON, and Markov Blanket Algorithm alongside with citation classifications, all features, only content features, bibliometric, and only the impact factor	It is feasible to accurately predict future citation counts with a mixture of content-based and bibliometric features using machine learning methods	Bio
Abramo et al. [2]	Linear regression, 8 years citation window to evaluate the Impact factor, early citations and compared the correlation of metrics (peer review, bibliometrics) with the success of scholarly publications	Both measures are not reliable and could be manipulative or biased after measuring the early impact of three years after the publication.	E,Ch
Li et al. [22]	Deep learning CNN prediction models, biblio-features.	Proposed methodology outperforms the state-of-the-art models and gives accurate prediction of future citations	M
Ruan et al. [29]	XGBoost, linear regression, four-layer Back Propagation (BP) neural network to predict the five-year citations of 49,834 papers, KNN, Random Forest, and Support Vector Regression.	Performance of the BP neural network is significantly better than the others, the accuracy of the model at predicting infrequently cited papers was higher than that for frequently cited ones. 5 features have effects ('citations in the first two years', 'first-cited age', 'paper length', 'month of the publication', and 'self-citations of journals')	Inf,Doc
Thelwall and Nevill [37]	Regression analysis of Altmetric.com data from November 2015 and Scopus citation counts from October 2017 for articles in 30 narrow fields	The main altmetric indicator of scholarly impact is Mendeley reader counts, journal impact factors can predict later citation counts better than Altmetric.com scores	Mut

CS=Computer Science, E=Engineering, Bio=Biomedical, Lib=Library, Inf=Information, S= Science, M=Mathematics, Re=Rehabilitation, PM=Physical Medicine, Me=Medical, L=Life,WoS=Web of Science, APS=Applied Physics Statistics, CPS=Computational Science, AM=Applied Mathematics, P=Physics,Ch=chemical, Mut=Mutiple fields. Doc=Documentation

Table 2Overview of the chosen Journals in our dataset from *Scopus*.

#	Journal	#Papers	SNIP	SJR	CSc	Publisher
1	Advanced Engineering Informatics	377	2.089	0.946	6.9	Elsevier
2	Computers and Education	1494	4.28	3.047	12.7	Elsevier
3	Engineering with Computers	254	2.014	0.663	7.2	Springer Nature
4	IEEE Transactions on Image Processing	2302	4.182	2.893	15.6	IEEE
5	IEEE Transactions on Information Forensics & Security	931	3.617	1.897	14.7	IEEE
6	Industrial Management and Data Systems	441	2.502	1.39	7.9	Emerald
7	Information Processing and Management	418	3.199	1.192	8.6	Elsevier
8	Journal of Informetrics	493	2.146	2.079	8.4	Elsevier
9	Journal of Machine Learning Research	1	3.147	2.219	9.3	MIT Press
10	Neural Networks	877	2.246	1.718	10.0	Elsevier
	Total	7,588				

#Papers from 2010 till 2015, CSc=CiteScore, SNIP=Source Normalized Impact per Paper, SJR=SCImago Journal Rank

4. Experiments

In this section, we describe how we elicited and analyzed the data to achieve our goal. Our experiment consists of four main phases: (1) Data collection, (2) Feature selection, (3) Classification predictive models, and (4) Evaluation of the models.

4.1. Data Collection

From Scopus digital library, we chose 10 computer science journals to extract their publications in a time period from 2010 to 2015 in order to give citation counts reasonable time to accumulate. The total number of extracted papers is 7,588. Table 2 displays the selected journals for our study and the number of papers extracted from each journal in addition to their properties displayed in the forms of the metrics that were used to choose them. We used two journal-related metrics in Scopus to help us decide which journals to include: CiteScore², Source Normalized Impact per Paper (SNIP)³, and SCImago Journal Rank (SJR)⁴.

The altmetrics were collected from PlumX⁵ tool which we chose because it is integrated with Scopus, and by using the available APIs, we were able to extract the needed features. Since we were interested in determining whether altmetrics can predict citations, we used citation counts for these articles as the target variables for our models. Using Scopus special APIs and by matching and merging the data for each article based on its DOI, citations were collected for each paper and our dataset was completed.

Our target variable is citation counts, and the features we selected are 14 altmetric features as listed and described in Table 3. The details of each feature are described on a separate webpage. Social media count⁶ are all interactions on social media platforms such as likes, shares on Facebook and Youtube, and tweets on Twitter. Subcategories of the Social media count are

²https://service.elsevier.com/app/answers/detail/a_id/14880/supporthub/scopus/

³https://service.elsevier.com/app/answers/detail/a_id/14884/supporthub/scopus/kw/SNIP/

⁴https://service.elsevier.com/app/answers/detail/a_id/14883/supporthub/scopus/kw/sjr/

⁵<https://plumanalytics.com/>

⁶<https://plumanalytics.com/learn/about-metrics/social-media-metrics/>

many, we chose tweet count and Facebook count. Mentions count⁷ is another PlumX category that includes blog posts, comments, reviews, and Wikipedia links about the publication from various resources such as Reddit, Slideshare, Vimeo, YouTube, and Github. The three most important subcategories of Mentions count are news, blog, and reference counts. Moving to the third category, Capture count⁸ which tracks users' actions like bookmarking, marking as favorite, reading, and exporting the paper. It also includes multiple subcategories such as reader count which gathers its data from CiteULike, Goodreads, Mendeley, and SSRN, and export/saves count. The last PlumX metrics category is Usage count⁹ that points out some usage statistics like links count, abstract or full-text view count and more.

The collected papers were ranked in descending order according to their citation counts and then categorized into 2 categories, the highly cited papers (HCPs) and the Low cited papers (LCPs) using the following process:

- Calculate the average of all citation counts
- Divide the papers according to their citations compared to the average into the following:
 - (HCPs) these papers have a number of citations that were greater than or equal to the average citation counts of the venue.
 - (LCPs) these papers have a number of citations less than the averaged citations of all papers of that venue.

The goal of this categorization of the papers into two groups is to characterize the various stages of paper growth, with HCPs being assigned to the successful ones. Despite the simplicity of this two-class classifier setting, the evaluation of which indications better forecast the future success of papers could be clearly captured based on it.

Table 3
Altmetrics features.

Index	Feature	Description
X ₀	Social media count	Number of times a paper has been mentioned or shared on any social network
X ₁	Tweet count	Number of times a paper has been mentioned in a tweet on Twitter
X ₂	FB count	Number of times a paper has been mentioned or shared in Facebook
X ₃	Mention count	Number of users who have mentioned a particular paper online
X ₄	News count	Number of times a paper has been mentioned in news outlets
X ₅	Blog count	Number of times a paper has been mentioned or featured in a blog post
X ₆	Reference count	Number of references of that particular paper
X ₇	Capture count	Number of times capturing the interest in the paper on the internet
X ₈	Reader count	Read is counted each time someone views the paper
X ₉	Export/Saves count	Represents the number of saves of the paper in external platform
X ₁₀	Usage count	Is a record of every action taken by all user
X ₁₁	Links click count	Number of clicks on the link of paper
X ₁₂	Links outs count	Number of link that will take you to the paper
X ₁₃	Full-text view	Number of times a paper has been fully viewed online

⁷<https://plumanalytics.com/learn/about-metrics/mention-metrics/>

⁸<https://plumanalytics.com/learn/about-metrics/capture-metrics/>

⁹<https://plumanalytics.com/learn/about-metrics/usage-metrics/>

4.2. Feature Selection

The term “feature selection” refers to the process of minimizing the number of input features that can be utilized to describe the interrelationships between them. It eliminates features that are redundant or useless. Irrelevant features give no valuable information about the data, whereas redundant features deliver no additional information than the currently selected features. In this paper, three different feature selection techniques were used to measure the importance of each feature on the dataset:

Variance Threshold (VAR) a fundamental baseline technique to feature selection is the VAR approach. It removes features with low variance or those whose variance is less than a particular threshold. The premise is easy to grasp. Calculate the standard deviation of each sample feature value and if the number is less than the threshold, filter and then eliminate. By default, all zero-variance characteristics are turned off. A variance of 0 shows that the sample feature’s value has remained unchanged.

$$\text{Var}[x] = p(1 - p)$$

Pearson’s Correlation (PC) Correlation-based Feature (CFS) is a well-known similarity measure that evaluates the correlation between features and classes, as well as between features and other features, to determine the significance of characteristics. In this paper, the importance of the features’ subset was determined by CFS using Pearson’s correlation equation. The covariance is $\text{cov}(X, Y)$. It can be used for binary classification and regression issues, with a range of $(-1, 1)$ from a negative to a positive correlation. It’s a fast statistic that ranks features according to their absolute correlation coefficient with the aim. Between a feature X and the target Y , the Pearson correlation coefficient is:

$$\rho_i = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma_y}$$

Mutual Information Gain (MI) is a metric for measuring how much information one random variable possesses about another. That is the mutual information between X and Y and may be thought of as a measure of X ’s amount of knowledge of Y (or Y ’s amount of knowledge of X). Therefore, it can be defined as:

$$I(X; Y) = H(X) - H(X|Y)$$

Where $I(X; Y)$ represents mutual information between X and Y , $H(X)$ represents entropy for X , and $H(X | Y)$ represents conditional entropy for X given Y .

4.3. Identification of the Prediction Algorithm

To evaluate the robustness of the feature subsets created using the three feature selection techniques, three machine learning algorithms based on classification were applied to the collected features. We tested the following supervised machine learning methods.

Decision Trees (DT) are trees that classify instances by sorting them based on feature values

[28]. Two entities, decision nodes, and leaves can be used to explain the tree. Each leaf in a decision tree indicates a value that the node might adopt, whereas each node represents a feature in an instance to be categorized. From the root to the leaf, a path is traced and sorted according to feature values. In this study, we have used 5 leaves.

Random Forest (RF) is a classifier made up of $h(x,k)$, $k=1,\dots$, where k is independently identically distributed random vectors, and each tree votes for the most popular class at input x with a single unit vote [9]. The RF classifier in this paper is made up of eight trees, each of which was developed using the classification and regression tree (CART) technique. Each case of a fresh dataset is handed down to each of the eight trees in order to categorize it. The forest picks the class with the most votes out of eight to be the case's final class label.

Support Vector Machines (SVM) is a technique, a sparse kernel decision machine that builds its learning model without calculating posterior probabilities. This is a relatively new supervised machine learning method. According to Gonzalez-Abril et al. [19], SVM conducts classification by creating an N -dimensional hyperplane that best divides the data into two groups. It has been demonstrated that increasing the margin and establishing the maximum feasible distance between the separating hyperplane and the instances on either side reduces the predicted generalization error.

4.4. Evaluation of Classification Models

We have three classifiers to select from to answer a specific classification issue, therefore, we need to assess the quality of each (prediction accuracy). To achieve that, we use a confusion matrix that describes the number of correctly and incorrectly predicted examples by the classification model. Table 4 depicts the binary classification problem's confusion matrix, which is a particular contingency table with two dimensions: actual and predicted. Each metric is a critical indicator of how well a model performed in relation to a set of criteria. The percentage of valid predictions correctly categorized by the model is known as model accuracy (Eq. 1):

$$Accuracy(Acc) = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Accuracy is the fraction of positive results predicted by the model that is really positive (Eq. 2):

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

Model recall is the proportion of relevant outcomes retrieved (Eq. 3):

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

5. Results and Discussion

The first step was selecting the most promising features using the three different feature selection methods, namely Variance Threshold (VAR), Pearson's Correlation (PC), and Mutual Information

Table 4

Definition of a confusion matrix.

		Predicted Class	
		Posiitive	Negative
Actual class	Positive	True Positive(TP)	False Negative(NP)
	Negative	False Positive(FP)	True Negative(TN)

(MI). The feature selection was done on the entire dataset. Our results show that there are 9 features that are considered the most significant to reflect the original dataset among the 14 features collected for the prediction task in Table 3. The outcome for the selected feature subset for each feature selection technique is shown in Table 5.

The indices Social count(X_0), Tweets(X_1), News(X_4), Capture(X_7), and Text view(X_{13}) appear in all three feature subsets among the nine features in Table 5, indicating that these five features constitute the dataset's fundamental characteristics and are the most important representative of the original dataset. That is, these five features are the ones that play the most important roles in deciding which papers will become highly cited. Indices Social count(X_0), Tweets(X_1) show how many times a paper has been discussed in a certain term and shared on the social network, this indicates how social media metrics help research dissemination. News(X_4) represents the number of times a paper is referenced in the news media. Moreover, Capture(X_7) captures the interest in the publication on the internet overall. Another useful feature is Text view(X_{13}), which is the number of times a publication has been seen in full detail.

The second step was splitting our dataset into test and training sets. We sample our training set while holding out 30% of the data for testing (evaluating) our classifier. This method can approximate how well our model will perform on new data.

The performance of these features in predicting future highly cited papers was then tested using the three classification models mentioned previously, Decision tree (DT), Random Forest (RF), and Support Vector Machines (SVM). A code project using Python and its libraries like SKlearn and numpy was developed to test these models and their performance later using the performance measures mentioned earlier. Table 6 illustrates the final classification performance of each feature selection method outcome under each of the three classifiers and the average classification accuracy (Acc), precision (Prc), and recall (Rcl) are shown in the last row. Obviously, each classifier has a significant classification performance for each of the feature subsets. The feature subset picked by Pearson's correlation (PC) and Mutual information (MI) has obtained the best precision with 0.97 respectively trained by Random Forest and in terms of precision, all three feature selection techniques have the same number with 0.96. Whereas the Variance threshold (VAR) has a maximum recall of 0.99, which was evaluated using Random Forest. Regardless of the classification model or feature selection approach, the average classification

Table 5

Feature results.

Selection	Feature subset
VAR	Social media(X_0),Tweets(X_1),News(X_4),Capture(X_7),Reader(X_8),Export(X_9),Usage(X_{10}),Links out(X_{12}),Text view(X_{13})
PC	Social media(X_0),Tweets(X_1),News(X_4),Blog(X_5),Capture(X_7),Export(X_9),Usage(X_{10}),Links out(X_{12}),Text view(X_{13})
MI	Social media(X_0),Tweets(X_1),News(X_4),Blog(X_5),Reference(X_6),Capture(X_7),Reader(X_8),Links clicks(X_{11}),Text view(X_{13})

Table 6
Classification Model Performance

	VAR			PC			MI		
Model	Acc	Prc	Rcl	Acc	Prc	Rcl	Acc	Prc	Rcl
DT	0.88	0.86	0.91	0.87	0.81	0.86	0.87	0.91	0.84
RF	0.97	0.96	0.99	0.97	0.96	0.98	0.97	0.96	0.98
SVM	0.93	0.95	0.91	0.92	0.94	0.9	0.92	0.94	0.9
Average	0.93	0.92	0.94	0.92	0.90	0.91	0.92	0.94	0.90

accuracy is equal to or more than 0.9. Although there is little variation in accuracies, the findings show that the features derived by the three feature selection approaches are stable and helpful to classify and forecast future highly cited papers. Furthermore, the results reveal that Random Forest, in particular, fared best, compared to the Decision Tree and Support Vector Machines.

The current study's drawback is that it only looked at 10 journals in the field of computer science. The findings from this small corpus would not apply to papers in other fields. Additionally, the current study is exclusively based on PlumX altmetrics correlated with Scopus and we narrowed our attention to only three machine-learning algorithms. Other algorithms, such as neural networks, and XGBoost might be investigated in the future. However, the results serve as a point of reference for future evaluations of prediction-related studies. We provide the dataset along with one prediction model code as an example for further analysis¹⁰.

6. Conclusion

In this paper, we build several experiments based on previous research that investigated metrics and their potential power to predict citation counts. Focusing on the computer science domain and aiming to find the most promising formula of altmetrics to predict the future success of a paper measured in the number of citations, we first performed several feature selection techniques to choose the most important feature subset that better represents the original dataset. An extensive dataset comprising papers from 10 computer engineering journals (7,588) was collected, altmetrics and citation counts for each paper were extracted. Furthermore, altmetrics were evaluated using a feature space with 14 feature indices to determine the most promising dataset using Variance threshold, Pearson's correlation, and Mutual information, and later the classification performance of the feature subsets was verified using three types of classifiers: Decision tree, Random forest, and Support vector machines. Finally, we evaluated these prediction models and compared their accuracy. The results show that Random forest surpasses the other classification methods and we conclude that altmetrics are a valuable predictor for highly cited papers, specifically these five altmetrics features: social media count, tweets, news, capture, reader count, and text view.

References

- [1] Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A., Arshad, H., 2018. State-of-the-art in artificial neural network applications: A survey. *Heliyon* .

¹⁰<https://doi.org/10.5281/zenodo.7777785>

- [2] Abramo, G., D'Angelo, C., Felici, G., 2019. Predicting publication long-term impact through a combination of early citations and journal impact factor. *Journal of Informetrics* .
- [3] Acuna, D.E., Allesina, S., Kording, K.P., 2012. Future impact: Predicting scientific success. *Nature* .
- [4] Akella, A.P., Alhoori, H., Kondamudi, P.R., Freeman, C., Zhou, H., 2021. Early indicators of scientific impact: Predicting citations with altmetrics. *Journal of Informetrics* .
- [5] Aksnes, D., Langfeldt, L., Wouters, P., 2019. Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open* .
- [6] Alchokr, R., Krüger, J., Shakeel, Y., Saake, G., Leich, T., 2022. Peer-reviewing and submission dynamics around top software-engineering venues: A juniors' perspective, in: *International Conference on Evaluation and Assessment in Software Engineering*.
- [7] Bai, X., Zhang, F., Lee, I., 2019. Predicting the citations of scholarly paper. *Journal of Informetrics* .
- [8] Bornmann, L., Leydesdorff, L., Wang, J., 2014. How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *Journal of Informetrics* .
- [9] Breiman, L., 2001. Random forests. *Machine Learning* .
- [10] Carlsson, H., 2009. Allocation of research funds using bibliometric indicators – asset and challenge to swedish higher education sector.
- [11] Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: *International Conference on Knowledge Discovery and Data Mining*.
- [12] Copiello, S., 2020. Other than detecting impact in advance, alternative metrics could act as early warning signs of retractions: tentative findings of a study into the papers retracted by plos one. *Scientometrics* .
- [13] Daud, A., Ahmad, M., Malik, M., Che, D., 2014. Using machine learning techniques for rising star prediction in co-author network. *Scientometrics* .
- [14] Edgar, T.W., Manz, D.O., 2017. *Machine Learning*. Syngress.
- [15] Eysenbach, G., 2011. Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research* .
- [16] Fan, J., Chen, M., Luo, J., Yang, S., Shi, J., Yao, Q., Zhang, X., Du, S., Qu, H., Cheng, Y., Ma, S., Zhang, M., Xu, X., Wang, Q., Zhan, S., 2021. The prediction of asymptomatic carotid atherosclerosis with electronic health records: A comparative study of six machine learning models. *BMC Medical Informatics and Decision Making* .
- [17] Fu, L., Aliferis, C., 2008. Models for predicting and explaining citation count of biomedical articles. *AMIA Symposium* .
- [18] Galligan, F., Dyas-Correia, S., 2013. Altmetrics: Rethinking the way we measure. *Serials Review* .
- [19] Gonzalez-Abril, L., Angulo, C., Velasco-Morente, F., Català, A., 2005. Unified dual for bi-class SVM approaches. *Pattern Recognition* .
- [20] Hassan, S.U., Aljohani, N., Idrees, N., Sarwar, R., Nawaz, R., Martínez-Cámara, E., Ventura, S., Herrera, F., 2020. Predicting literature's early impact with sentiment analysis in twitter. *Knowledge-Based Systems* .
- [21] Holden, G., Rosenberg, G., Barker, K., 2005. Tracing thought through time and space: A

- selective review of bibliometrics in social work. *Social Work in Health Care* .
- [22] Li, M., Xu, J., Ge, B., Liu, J., Jiang, J., Zhao, Q., 2019. A deep learning methodology for citation count prediction with large-scale biblio-features.
- [23] Lutz, B., 2014. Do altmetrics point to the broader impact of research? an overview of benefits and disadvantages of altmetrics. *Journal of Informetrics* .
- [24] Nuzzolese, A.G., Ciancarini, P., Gangemi, A., Peroni, S., Poggi, F., Presutti, V., 2019. Do altmetrics work for assessing research quality? *Scientometrics* .
- [25] Patro, B., Aggarwal, A., 2011. How honest is the h-index in measuring individual research output? *Journal of postgraduate medicine* .
- [26] Penner, O., Pan, R.K., Petersen, A.M., Kaski, K., Fortunato, S., 2013. On the predictability of future impact in science. *Scientific reports* 3, 3052.
- [27] Poggi, F., Ciancarini, P., Gangemi, A., Nuzzolese, A.G., Peroni, S., Presutti, V., 2019. Predicting the results of evaluation procedures of academics. *PeerJ Computer Science* .
- [28] Quinlan, J.R., 1986. Induction of decision trees. *Machine Learning* .
- [29] Ruan, X., Zhu, Y., Li, J., Cheng, Y., 2020. Predicting the citation counts of individual papers via a BP neural network. *Journal of Informetrics* .
- [30] Shakeel, Y., Alchokr, R., Krüger, J., Leich, T., Saake, G., 2022a. Altmetrics and citation counts: An empirical analysis of the computer science domain, in: *Joint Conference on Digital Libraries*.
- [31] Shakeel, Y., Alchokr, R., Krüger, J., Leich, T., Saake, G., 2022b. Are altmetrics useful for assessing scientific impact? a survey, in: *International Conference on Management of Digital EcoSystems*.
- [32] Shakeel, Y., Alchokr, R., Krüger, J., Leich, T., Saake, G., 2022c. Incorporating altmetrics to support selection and assessment of publications during literature analyses, in: *International Conference on Evaluation and Assessment in Software Engineering*.
- [33] Shakeel, Y., Alchokr, R., Krüger, J., Saake, G., Leich, T., 2021. Are altmetrics proxies or complements to citations for assessing impact in computer science?, in: *Joint Conference on Digital Libraries*.
- [34] Siler, K., Lee, K., Bero, L., 2015. Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences* .
- [35] Stegehuis, C., Litvak, N., Waltman, L., 2015. Predicting the long-term citation impact of recent publications. *Journal of Informetrics* .
- [36] Thelwall, M., 2020. The pros and cons of the use of altmetrics in research assessment. *Scholarly Assessment Reports* .
- [37] Thelwall, M., Nevill, T., 2018. Could scientists use altmetric.com scores to predict longer term citation counts? *Journal of Informetrics* .
- [38] Wang, D., Song, C., Barabási, A.L., 2013. Quantifying long-term scientific impact. *Science* .
- [39] Wang, M., Wang, Z., Chen, G., 2019. Which can better predict the future success of articles? Bibliometric indices or alternative metrics. *Scientometrics* .
- [40] Yu, T., Yu, G., Li, P.Y., Wang, L., 2014. Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics* 101, 1233–1252.