Information Research - Vol. 31 No. 1 (2026)

Citation count prediction based on Google Scholar profiles and Clarivate's journal citation reports

Mahdi Bahaghighat, Leila Akbari, Majid Ghasemi, and Qin Xin

DOI: https://doi.org/10.47989/ir31141005

Abstract

Introduction. Citation count prediction (CCP) models are vital for assessing research impact, yet existing approaches suffer from critical limitations. Prior studies often rely on restricted datasets (e.g., journal metrics alone) or fail to account for the multidimensional factors influencing citations, leading to suboptimal accuracy.

Method. We propose an accurate CCP regression model for Computer Science and Electrical Engineering disciplines found on twenty three novel features extracted from public data in Google Scholar profiles and the Journal Citation Reports (JCR) annual report by splitting features into four datasets: Author information database (AI DB), journal information database (JI DB), paper information database (PI DB), and finally author & paper & journal information database (APJ DB).

Analysis. Our evaluation employed Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2) to assess model performance. Dimensionality reduction techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) were also applied, and their effect on CCP was assessed.

Results. We identified that paper-level features (PI DB) were significantly more predictive than author or journal attributes, resolving a key debate in CCP research.

Conclusions. This study enhances CCP research by introducing scalable, publicly available features, demonstrating the superiority of paper-level attributes through empirical evidence, and identifying Nu-SVR as the most effective algorithm for accurate and interpretable citation prediction, supporting researchers, institutions, and policymakers in assessing research impact.

Introduction

Citation is considered in scientometrics as the reference of one published document in another document. It also is a degree to evaluate the prominence and measure of research outputs because it presents how usually a specific work is referenced in alternative scholarly literature. A high citation count often indicates that a work has significantly impacted its field, influencing subsequent research and scholarship (Blümel & Schniedermann, 2020; Broadus, 1987; Moed, 2006; Khokhlov, 2020). So it can act as an indicator of research quality; works with high citation counts are often thought to be more valuable or credible than less cited peer-reviewed literature (Bornmann & Daniel, 2008; Butler & Visser, 2006). Citation counts can also provide to spot trends in research over time, indicating areas of growing interest, and shifts in the scientific landscape. Besides, citation counts play a key role in resource allocation decisions, as funding agencies and institutions also rely on these metrics to decide which research areas or researchers are worthy of support. Overall, citation counts play a vital role in academic communication, evaluation, and the advancement of knowledge in various fields (Abramo et al., 2023; Belikov & Belikov, 2015; Cao et al., 2016; Durieux & Gevenois, 2010; Gao et al., 2024; Groos & Pritchard, 1969; Sohrabi & Iraj, 2017; Yu et al., 2014).

On the other side, several famous bibliometrics measures, including Impact Factor (IF) and h-index, are also based on citations from publications and journals. In the early 1960s, the IF was introduced by the Institute of Scientific Information (ISI) (Garfield, 2006). The IF is determined by the number of citations it receives throughout the year. The IF is calculated by dividing the current number of citations by the number of articles published in the last two years. As a result, the performance of a journal can be directly estimated based on its IF (Amin & Mabe, 2003; Bornmann & Daniel, 2009; Braun et al., 2006; Durieux & Gevenois, 2010; Hirsch, 2010; Lundberg, 2006). It has been criticized for not taking into account the diversity of individual researchers among different fields of study and for being inaccurate in terms of its purpose. In addition, researchers were given an index called the h-index in order to measure their objective function. To determine an author's h-index, one must determine how many of his or her articles have been cited directly by other researchers at least a similar number of times (Braun et al., 2006; Fassin, 2020; Hirsch, 2010; Khurana & Sharma, 2022). In addition, Q1 to Q4 quartile rankings for journals are provided by two key sources: Clarivate Analytics and Elsevier. Clarivate's Journal Citation Reports (JCR) ranks journals based on their Impact Factor (IF) through its Web of Science platform, assigning quartiles from Q1 (top 25%) to Q4 (bottom 25%). Elsevier's Scopus platform uses the SCImago Journal Rank (SJR), which evaluates journals based on citation impact and assigns similar quartiles. Both systems are widely recognized and used for academic journal rankings globally, guiding researchers in assessing and selecting journals for publishing their work (Almas et al., 2021; González-Betancor & Dorta-González, 2017; Kosyakov & Pislyakov, 2024; Moussa, 2023; Okagbue et al., 2020; Okagbue et al., 2021; Teixeira da Silva, 2020; Torres-Salinas et al., 2022).

Gao et al., 2024 leveraged multi-layer academic networks to improve citation count predictions by fusing different relationship types between publications. The authors show how taking this multilayer perspective of academic relations can enhance precision. Their model offers the ability to capture complexity in citation dynamics. Nevertheless, the intricacy of a multi-layered network could impose difficulties in model interpretability, hindering researchers from interpreting the mechanisms that power predictions. Also, the model performance may depend on the quality of the network that it is trained on, which may differ greatly between different sectors.

The process of peer reviewing is widely accepted as a means of evaluating papers (Li et al., 2019). It is important for a reviewer to evaluate the originality, creativity, contribution, integrity, and readability of a paper. Since peer-reviewing data includes the assessment comments of related experts, it may also be possible to predict the future influence of a paper. The algorithm of a comprehensive review paper has made it possible to obtain peer-reviewed data for the Citation

count prediction (CCP) task. It is apparent in their comments that they focus on issues that are not directly related to the paper's main contribution. Reviewers may include reminders regarding formatting issues in their reviews. Several people may be reviewing articles at the same time, resulting in differing opinions. Therefore, when determining the impact of a paper, the coverage and divergence of review comments would both be considered (Li et al., 2019).

While citation count prediction is a relatively well-explored area, many existing studies did not comprehensively consider the multitude of factors influencing citations, particularly in specialized fields like Computer Science and Electrical Engineering (Aksnes et al., 2019; Baas et al., 2020; Enduri et al., 2022; Furman & Teodoridis, 2020; He et al., n.d.; Zhang et al., 2025; Hutchins et al., 2016; Okagbue et al., 2020).

Citation patterns can vary across disciplines, so our work aimed at citation count predictions of academic papers for computer science and electrical engineering disciplines. Due to the fact that both Google Scholar and JCR are well-respected sources in the academic community and public availability of data, we suggested mainly gathering raw data from Google Scholar Profiles (GSPs) and public JCR annual reports. The GSP offers a wide range of citation information, such as Citation, h-index, and i10-index, while the JCR reports provide metrics like IF and journal rankings (Q1, Q2, Q3, Q4). Combining these two main data sources can enrich our feature set and help capture more nuanced aspects of citation behavior, making our model more accurate. Raw data like citation counts, h-index, or publication year are useful but limited in their predictive power. Citations are influenced by a variety of factors, and raw data may not fully capture non-linear relationships between these factors. Creating new features can improve the performance of the citation count prediction model by uncovering hidden patterns and relationships that raw data alone cannot reveal. As a result, our approach includes twenty three unique and novel features, offering a fundamental understanding of the factors that may impact citation count. The citation count was then estimated using a number of robust regression techniques, including SVR, Nu-SVR, Linear SVR, K-Nearest Neighbors (KNN), Decision Tree (DT), Bayesian Ridge, and SGD Regressor. We also assessed our method using several performance measures including the mean absolute percentage error (MAPE), Mean Square Error (MSE), Mean Absolute Error (MAE), and R-squared. The study also utilized PCA and t-SNE for dimensionality reduction and examined their impacts on CCP.

The rest of this paper is organized as follows: Section two discusses related works. In section three, our methodology is introduced. Simulation results are presented in section four, and section five is the conclusion.

Related work

Citation count prediction (CCP) has been approached from multiple perspectives, including network-based modeling, textual analysis, trend forecasting, and deep learning. While existing methods offer valuable insights, they often focus on isolated aspects of the problem, leaving room for a more holistic and generalizable approach (Bai et al., 2025; He et al.,2025; Zhang et al, 2025, Zhu et al.,2025).

Early work by (Pobiedina & Ichise, 2016) framed CCP as a link prediction problem, leveraging citation networks to model future citations. While this approach captures structural dependencies, it overlooks critical external factors such as author reputation and research competitiveness. Similarly, (Wang et al., 2023) introduced AGSTA-NET, a spatio-temporal fusion model that improves dynamic citation network analysis. However, its computational complexity and reliance on heterogeneous network data may limit scalability. These studies highlight the potential of network-based methods but also reveal their dependence on well-structured citation data, which may not always be available

Recent work has explored the role of textual features in CCP. For example, (Li et al., 2019) incorporated peer review text into a neural network model, demonstrating that qualitative feedback can enhance prediction accuracy. However, their reliance on peer review data—which varies widely across disciplines—poses a generalizability challenge. Similarly, (Sohrabi & Iraj, 2017) focused on keyword frequency, showing that strategic keyword use can improve visibility. Yet, their model neglects broader contextual factors, such as journal prestige or research impact. Baba et al. (2019) extended textual analysis to paper abstracts but did not account for external citation influences. These studies suggest that while textual features are valuable, they must be integrated with other predictive factors for robust performance.

Historical citation trends have also been used for prediction. (Li et al., 2015) demonstrated that temporal patterns improve out-of-time forecasts, but their model struggles with disruptive research that defies conventional citation trends. Meanwhile, (Abrishami & Aliakbary, 2019) applied deep learning, achieving superior accuracy over traditional methods. However, their approach requires large datasets and risks overfitting, limiting applicability in low-data scenarios.

Existing CCP methods face key limitations: network models ignore external factors; textual approaches lackgeneralizability; trend-based methods fail with disruptive research; and deep learning requires excessive data. Most critically, no unified framework integrates multi-modal data while ensuring efficiency and interpretability (Nguyen et al., 2025; Zafar et al., 2024).

Methodology

In this paper, we aim to predict the citation count of an article. An article's citation count is a suitable determinant for the impact assessment. For this purpose, we have created and developed four datasets that are called Author Information Database (AI DB), Journal Information Database (JI DB), Paper Information Database (PI DB), and finally Author & Paper & Journal Information Database (APJ DB). Figure 1 depicts the proposed algorithm. The initial preprocessing step normalizes the data. Then, we can predict the citation count of a published paper using several robust regression algorithms such as K-Nearest Neighbors (KNN), Decision Trees (DT), and Support Vector Regression (SVR), and Bayesian Regression methods, along with some of the most important performance metrics in regression problems such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2) for evaluation of the achieved results. To improve interpretability, we also assessed the influence of dimensionality reduction techniques (e.g., t-SNE and PCA) on the results and discussed their comparative effects.

From data to information: Creating some comprehensive database

The data collection process was conducted in two phases. In the first phase, at the end of December 2022, all input features were gathered. The second phase, at the end of 2023, involved collecting the output data (predictions). Data sources included Google Scholar profiles, the Journal Citation Reports (JCR) annual report, and SCImago (to obtain the SJR metric). The dataset only focuses on papers in Computer Science (CS) and Electrical Engineering (EE). Additionally, we developed a specialized dataset called AoI2WoS (Bahaghighat et al., 2024; Jahani rad et al., 2024), which establishes a connection between Areas of Interest (AoI) in GSP and Web of Science (WoS) scientific fields. This dataset was used to evaluate whether a given GSP is related to Computer Science (CS) or Electrical Engineering (EE), allowing us to filter out irrelevant profiles and focus on approximately 2,000 papers from randomly selected authors in these fields.

To create the AI DB, the author's scholarly background is examined found on ten suggested attributes. AI DB includes the information of the authors such as N_p, $N_p^{C=0}$, $C_{max}^{N_p}$, S_{max}^{10}, $citation$, $h-index^{Author}$, $i10-index$, Y_{FP}, Y_{LP}, and T_W^{Author} which can be seen with more details in Table 1. The second dataset is the JI DB. According to Table 2, we have gathered numerous critical information

such as Impact Factor (IF), h-index, SJR, Q1, Q2, Q3, Q4, and Q (Best quartile among all disciplines) in the JI DB. The third dataset is called the PI DB, in which some features of published papers were defined. According to Table 3, some attributes such as T_W^{Paper} (availability of the paper from publication year to current year), $C_m^{publication-year}$ (citation of the manuscript in publication year), $C_m^{last-year}$ (citation of the manuscript in the last year), and N^A (Number of Authors). Finally, the APJ DB has been constructed based on all information available in all three mentioned datasets (including all twenty-three defined features).

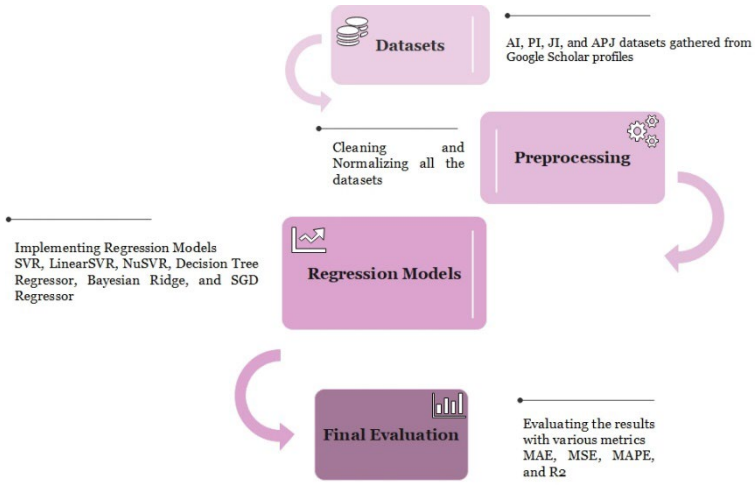

Figure 1. An illustration of the proposed citation count prediction algorithm (CCP).

Row	Attributes	Comments
1	N_p	Total number of publications for the author
2	$N_p^{C=0}$	Total number of publications without citations for the author
3	$C_{max}^{N_p}$	The most citation for an author: $max\ (C_i)$,for $i = 1$ to N_p
4	S_{max}^{10}	Sum of top ten citation for an author: $S_{max}^{10} = \sum_{i=1}^{i=10} C_{max}^{ith}$
5	$citation$	Total citations for an author
6	$h-index^{Author}$	h-index for an author
7	$i10-index$	i10-index for an author
8	Y_{FP}	First publication year: Author's first publication (in year)
9	Y_{LP}	Last publication year: Author's last publication (in year)
10	T_W^{Author}	$T_W^{Author} = Y_{LP} - Y_{FP}$

Table 1. Proposed attributes in author information dataset (AI DB).

Figure 2 shows an example of an author's Google Scholar profile. In addition, some characteristics of the proposed dataset can be seen in Figure 3, and Figure 4.

Row	Attributes	Comments
1	IF	The Impact Factor of a journal
2	Q1	Top 25% of journals (if a journal is Q1 then Q1=1, Q2=Q3=Q4=0)
3	Q2	25% to 50% of journals (if a journal is Q2 then Q2=1, Q1=Q3=Q4=0)
4	Q3	50% to 75% of journals (if a journal is Q3 then Q3=1, Q1=Q2=Q4=0)
5	Q4	75% to 100% of journals (if a journal is Q4 then Q4=1, Q1=Q2=Q3=0)
6	Q	Q is derived from Q1 to Q4. It is equal to 1.00 for Q_1; 0.75 for Q_2; 0.50 for Q_3; 0.25 for Q_4 according to the best quartile among all disciplines
7	SJR	The SCImago Journal Rank

Table 2. Proposed attributes in journal information dataset (JI DB).

Row	Attributes	Comments
1	T_W^{Paper}	It shows how many years the paper is available online
2	$C_m^{Publication-year}$	The total citations of the paper in the published year
3	$C_m^{Last-year}$	The total citations of the paper in the last year
4	N^A	Number of authors in a paper
5	OAA	Open Access Article

Table 3. Proposed attributes in paper information dataset (PI DB).

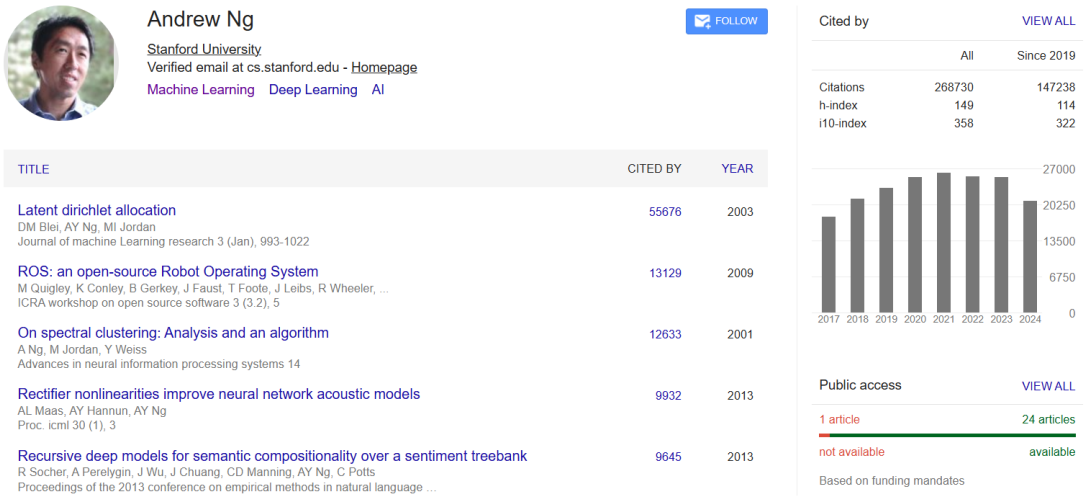

Figure 2. The author information in its Google Scholar profile.

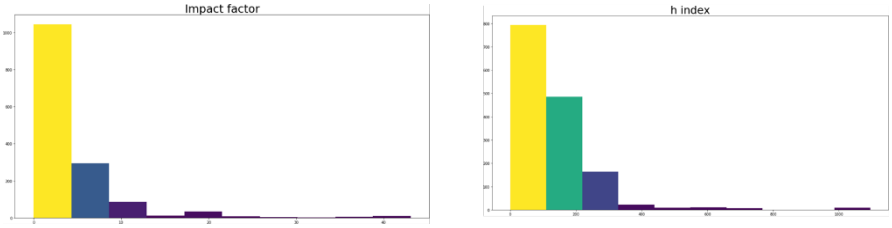

Figure 3. The histograms of some features in the JI DB dataset

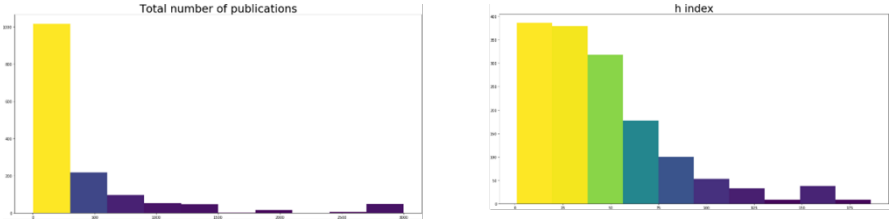

Figure 4. The histograms of some features in the AI DB dataset.

Regression models

In statistical approaches, Regression Analysis (RA) is considered a set of statistical procedures to estimate the relationships between an output (dependent variable) and one or more inputs (independent variables). As a statistical method, regression describes the relationship between two or more inputs and outputs (Fox, 2015; Rostami et al., 2021). In this paper, we deploy several strong regression models to predict the number of citations of a paper.

SVR and linearSVR

Machine learning (ML) is a sub-field of Artificial Intelligence (AI) that enables systems to learn automatically, and as opposed to being explicitly programmed, they can improve their decision-making abilities by acquiring experience (Chen et al., 2024). Support Vector Regression (SVR) (Hearst et al., 1998; Smola & Schölkopf, 2004) distinguishes itself by employing the Structural Risk Minimization (SRM) principle, a foundation rooted in statistical learning theory. SRM's core objective is to craft a hypothesis (h) that minimizes the true error when applied to unseen and randomly sampled testing data. Notably, SVR excels in handling outliers, a critical advantage in practical applications. In general, SVR estimation functions have the following form (Basak et al., 2007; Smola & Schölkopf, 2004):

$$f(x) = (w \cdot \phi(x)) + b \qquad (1)$$

Where $b \subset R$ and ϕ indicate a nonlinear conversion from R^n (the real coordinate space or real coordinate n-space, of dimension n) to high-dimensional space, the aim is to indicate the value of w. In order to compute the value of x, it is necessary to minimize the regression risk. Minimizing the regression risk determines the values of x.

$$R_{reg}(f) = C \sum_{i=0}^{l} \Gamma\left(f(x_i) - y_i\right) + \frac{1}{2} \parallel w \parallel^2 \qquad (2)$$

The cost function is $\Gamma(.)$. In Support Vector Regression (SVR), the cost function aims to minimize the discrepancy between the model's output and the actual values corresponding to the training data. C is a constant value, and vector w can be calculated as below:

$$w = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)\, \phi(x_i) \qquad (3)$$

Using substituting Eq. (3) into Eq. (1), the general equation can be revised as the Eq. (4):

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \left(\phi(x_i) \cdot \phi(x)\right) + b = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)\, k(x_i.x) + b \qquad (4)$$

In Eq. (4), the function was replaced with dot product $k(x_i, x)$, that function $k(x_i, x)$ familiar as the kernel function. In a high-dimensional feature space, evaluate the dot function based on low-dimensional input data without understanding how the transformation is performed. There is a condition of Mercer that all kernel functions must satisfy, which is equivalent to the inner product of some feature space. For regression, the Radial Basis Function (RBF) is used as the standard kernel. RBF kernels are presented in the following equation.

$$k(x_i.x) = exp\{-Y \parallel x - x_i \parallel^2\} \qquad (5)$$

A few standard kernels in SVR are Linear ($x \times y$), Polynomial ($[(x \times x_i) + 1]^d$), Radial Basis Function ($exp\{-\gamma \parallel x - x_i \parallel^2\}$) are shown in Table 4.

Kernel	Function
Linear	Simple, faster, and lower accuracy for nonlinear data
Polynomial	Fast and more flexible
RBF	More flexible and higher accuracy

Table 4. Common kernel function (Basak et al., 2007; Smola & Schölkopf, 2004).

Nu-SVR

In Nu-SVR, the inclusion of the epsilon parameter allows for the control of the number of support vectors retained in the solution. The introduction of the nu parameter further provides a mechanism to manage support vectors by specifying the proportion of these vectors relative to the total number of samples in the dataset. Given a sample pair of the dataset, including input-output $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, the Nu-SVR method is used to approximate a nonlinear relationship. It is used to minimize overfitting by getting as close as possible to the target function (Bhatt et al., 2012; Smola & Schölkopf, 2004). The types of kernels that could be used are a polynomial function, Radial Basis Function (RBF), a sigmoid function, and a linear function.

KNN

The k-Nearest Neighbors (kNN) method (Cover & Hart, 1967)is widely adopted in data mining and statistics for its simplicity and notable classification performance (Cunningham & Delany, 2021; Halder et al., 2024). Despite its ease of implementation, the kNN method has demonstrated significant classification prowess and has been shown to approximate the error rate of Bayes optimization under mild conditions. Its versatility extends to various applications, including regression, classification, and missing value imputation. However, the efficacy of the kNN method is contingent upon factors like the choice of the k value and the selection of distance measures. Addressing these considerations has led to the development of numerous machine learning techniques aimed at optimizing the performance of the kNN method. In order to solve classification problems, KNN is a highly beneficial approach (Halder et al., 2024; Sabry, 2023; Song et al., 2017). The unique property was calculated where no explicit step is required in the training phase (other than the capacity of the training database). Based on the data in the testing dataset, kNN is used to evaluate the answer to x_t as a weighted mean of the responses of the k nearest training points $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$ the neighborhood of x_t. This function can calculate how near each training data point x_i is to the testing data point x_t using the weighted Euclidean distance, described as (Halder et al., 2024; Sabry, 2023):

$$d(x_t.x_i) = \sqrt{\sum_{n=1}^{N} W_n\,(x_{t.n}.x_{i.n})^2} \qquad (6)$$

Then, we apply the kernel of the regression and calculate the following approximation of the response of x_t (Yao & Ruzzo, 2006):

$$\hat{f}(X_t) = \frac{\sum_{i=1}^{k} \sigma\left(X_t.x_{(i)}\right) f(X_i)}{\sum_{i=1}^{k} \sigma\left(X_t.x_{(i)}\right)} \qquad (7)$$

Decision tree

Decision rees are a widely used ML method for both classification and regression problems. It consists of a tree structure with internal nodes containing tests on features, branches representing the output of the tests, and leaf nodes containing the predicted output values when the path from the root to the leaf node has been established. Then, the data is split recursively on the feature that would result in the best split, often according to criteria like mean squared error or variance. Decision Trees can handle continuous and categorical variables, and in the case of continuous variables, the algorithm seeks a threshold in order to perform the split. Their simplicity and interpretability, as well as a potential to highlight relevant features, make them widely utilized for feature selection (Bishop, 2006).

Bayesian Regression

Bayesian Regression uses Bayesian statistics to estimate the parameters of the regression model. It uses some prior beliefs regarding the parameters, which are modified using observed data, to form posterior distribution through the Bayes Theorem. Bayesian Regression, on the other hand, provides a posterior distribution over parameters, which is a more robust and flexible approach than classical regression methods. It is however, computationally expensive and requires the proper choice of prior distributions, which would add some subjective influence. The Bayesian Regression could give you better estimates, but it is not without its challenges. However, it is also widely used in fields such as finance, engineering, social sciences, etc. (Bishop, 2006; Pedregosa et al., 2011).

SGDRegressor

It is basically a linear regression in scikit-learn with a Stochastic Gradient Descent (SGD) optimiser to learn the optimal set of parameters for our model. In contrast to the previous update, traditional gradient descent updates weights after the whole dataset is processed, while SGD performs it for every single sample, which is therefore more computationally effective, particularly with large datasets. This model is perfect when your features are sparse and your data is high-dimensional (Pedregosa et al., 2011).

Dimension reduction

Essentially, Dimension reduction (DR) is one of those very prominent techniques you would often use if working on machine learning or data analysis projects, and it basically simplifies your data sets by reducing the number of variables to only those that are significant. PCA (Principal Component Analysis) is probably the most famous DR method, which converts the data into a new set of variables (principal components) that are all orthogonal (no correlation). PCA achieves dimensionality reduction with minimum data loss by choosing the best components. However, the effectiveness of PCA is closely related to the data/model you use. This can eliminate noise, too, but this sacrifices interpretability and might go with it, model performance. Hence comparison should be made between models with and without PCA to check the effect (Bishop, 2006; Pedregosa et al., 2011). In our analysis, we applied both t-SNE (t-Distributed Stochastic Neighbor Embedding), and PCA for dimensionality reduction. t-SNE is a nonlinear dimensionality reduction technique widely used for visualizing high-dimensional data. Unlike linear methods such as PCA,

it focuses on preserving local similarities between data points by modeling pairwise probabilities in both the original and reduced spaces. It employs a t-distribution in the low-dimensional space to mitigate crowding effects, making it particularly effective for revealing clusters or manifolds in complex datasets (e.g., images or biological data). However, t-SNE's results can be sensitive to hyperparameters (e.g., perplexity) and computationally intensive for large datasets (Skrodzki et al., 2024; Jung et al., 2024).

Performance metrics

Here we discuss methods for calculating errors in the proposed CCP model. In our research, we have used a variety of error measurement methods, such as Mean Absolute Error (MAE), Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), and R-squared (R2). An important loss function in regression analysis is the Mean Square Error (MSE) (Cameron & Windmeijer, 1997; Murphy, 1988 ;Wallisch et al., 2022; Willmott & Matsuura, 2005). The mean squared distance between the predicted and actual values is calculated using this loss function. Below is a description of how it is calculated:

$$MSE = \frac{\sum_{i=1}^{n}(prediction_i - true_i)^2}{n} \quad (8)$$

Other errors that have been used in this paper are MAE, MAPE, and R2:

$$MAE = \frac{\sum_{i=1}^{n} p\,rediction_i - true_i}{n} \quad (9)$$

$$MAPE = \frac{\sum_{i=1}^{n} \frac{true_{i_{prediction_i}}}{true_i}}{n} \quad (10)$$

$$R^2 = \frac{\sum_{i=1}^{n}(prediction - mtrue_i)}{\sum_{i=1}^{n}(true_i - mtrue_i)} \quad (11)$$

$$m_{true_{(i)}} = \frac{1}{n}\sum_{i=1}^{n} t\,rue_i \quad (12)$$

Results

In this paper, seven distinct regression methods, including SVR, Nu-SVR, Linear SVR, kNN, Decision Tree Regression, Bayesian Ridge, and SGD Regression, were trained to find the best Citation count prediction (CCP) solution. We also used several error methods for estimating the errors, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2), to measure the algorithm performance for the Author information database (AI DB), journal information database (JI DB), paper information database (PI DB), and finally author & paper & journal information database (APJ DB), separately. We also tested APJ DB and AI DB in two scenarios, with dimension reduction and without it. All of our simulations have been implemented using Python software and were run on a Lenovo device with a 2.5 GHz Intel Core i7 processor and 16 GB DDR4 RAM. The results obtained from the different methods are presented from Table 5 to Table 17.

In our experiment, Grid search that is a hyperparameter tuning technique was used to systematically work through multiple combinations of parameter values to determine which combination yields the best model performance.

Model	Best Parameters
SVR	{'C': 29.108, 'gamma': 0.00156, 'kernel': 'rbf'}
Nu-SVR	{'C': 122.519, 'gamma': 0.000535, 'kernel': 'rbf', 'nu': 1}
LinearSVR	{'C': 0.01672, 'epsilon': 0.01682, 'loss': 'epsilon_insensitive'}
KNeighbors Regressor	{'n_neighbors': 1}
Decision Tree Regressor	{'criterion': 'squared_error', 'max_depth': 24.42, 'splitter': 'random'}
Bayesian Ridge	{'alpha_1': 3.59e6, 'alpha_2': 10.0, 'lambda_1': 1e11, 'lambda_2': 21544.35, 'n_iter': 1}
SGD Regressor	{'alpha': 0.07508, 'epsilon': 12.07867, 'penalty': 'elasticnet'}

Table 5. Primary grid search results for suggested models.

In Table 5, max_depth=24.42 suggests a deep tree for DecisionTree Regressor, which may be harder to interpret than a shallower one. So, we could constrain max_depth to a smaller value about 5 during grid search. Furthermore, KNeighborsRegressor with n_neighbors=1 may highly prone to overfitting and lacks interpretability. As a result, a higher value at 5 was selected. LinearSVR and SGDRegressor use regularization (C, alpha). A low C=0.01672 (LinearSVR) suggests strong L2 regularization, which can simplify the model by shrinking coefficients. In addition, SGDRegressor uses penalty='elasticnet', which can promote sparsity, making feature importance clearer. For Kernel Choices in SVM Models Both SVR and Nu-SVR use the 'rbf' kernel, which is inherently less interpretable than a linear kernel. As a result, for considering interpretability, we could restrict the search to kernel='linear'. For Bayesian Ridge's Complexity, the high values for lambda_1=1e11 and alpha_1=3.59e6 suggest strong prior assumptions, but the model remains interpretable since it's a linear regression variant.

Table 6 presents the performance of various regression models on the APJ dataset without any dimensionality reduction. The best-performing models are SVR and Decision Tree Regression, both achieving an MAE of 0.16, MSE of 0.50, and R^2 of 0.48, indicating strong predictive accuracy and stability. Nu-SVR also performs well with the lowest MAPE (0.36), suggesting better relative error control. In contrast, K-Neighbors Regression and Bayesian Ridge show slightly higher errors, while Linear SVR and SGDRegression exhibit the weakest R^2 scores (0.35 and 0.43, respectively). Overall, non-linear models (SVR, Decision Tree) outperform linear ones, likely due to their ability to capture complex relationships in the data.

Regression models	MAE	MSE	MAPE	R2
SVR	0.17	0.50	0.50	0.48
Nu-SVR	0.16	0.56	0.36	0.43
Linear SVR	0.17	0.63	0.43	0.35
K-Neighbors Regression	0.20	0.55	0.64	0.43
Decision tree regression	0.16	0.50	0.52	0.48
Bayesian Ridge	0.19	0.54	0.61	0.44
SGDRegression	0.19	0.55	0.57	0.43

Table 6. Regression results for the APJ DB and without dimension reduction.

According Table 7, applying PCA-based dimension reduction generally worsened model performance, except for K-Neighbors Regression, which saw slight improvements (e.g., MSE dropping from 0.55 to 0.53). The degradation is most severe for Decision Tree Regression, where MSE nearly doubled (0.50 → 0.92) and R^2 collapsed to 0.05, because PCA's linear transformations disrupt the tree-based feature splits. Similarly, SVR and Nu-SVR suffered, possibly due to lost non-linear feature interactions. The improvement in K-Neighbors suggests PCA may have removed noise, aiding its distance-based computations. The overall decline implies that PCA either discarded informative features or failed to preserve structures critical for regression, highlighting that blind dimensionality reduction can harm performance unless the model benefits from noise removal (like kNN).

Regression models	MAE	MSE	MAPE	R2
SVR	0.20	0.55	0.67	0.44
Nu-SVR	0.19	0.64	0.53	0.34
Linear SVR	0.21	0.57	0.82	0.41
K-Neighbors Regression	0.18	0.53	0.60	0.45
Decision tree regression	0.26	0.92	1.12	0.05
Bayesian Ridge	0.20	0.56	0.72	0.42
SGDRegression	0.21	0.55	0.76	0.42

Table 7. Regression results for the APJ DB and with dimension reduction based on PCA.

Comparing both tables show that PCA does not universally improve model performance and even degrades results for certain algorithms. Given this outcome, this study further evaluates t-SNE (t-Distributed Stochastic Neighbor Embedding) as an alternative and compares results with PCA. The Table 8 presents MAE and MSE values for six regression models, with performance measured after applying PCA and t-SNE for dimensionality reduction. While t-SNE excels at visualization, PCA remains almost superior for regression tasks due to its stability, interpretability, and preservation of globally meaningful features. The results shows that t-SNE cannot outperform PCA in regression settings when it was applied to APJ DB.

Regression Model	Metric	PCA	t-SNE
SVR	MAE	0.20	0.26
SVR	MSE	0.55	0.85
Nu-SVR	MAE	0.19	0.23
Nu-SVR	MSE	0.64	0.94
Linear SVR	MAE	0.21	0.28
Linear SVR	MSE	0.57	1.10
K-Neighbors	MAE	0.18	0.33
K-Neighbors	MSE	0.53	0.52
Decision Tree	MAE	0.26	0.25
Decision Tree	MSE	0.92	1.15
Bayesian Ridge	MAE	0.20	0.32
Bayesian Ridge	MSE	0.56	0.72
SGDRegression	MAE	0.21	0.25
SGDRegression	MSE	0.55	0.99

Table 8. Regression results for APJ DB: PCA vs. t-SNE dimensionality reduction.

Table 9 presents the regression results for the AI DB without dimension reduction. The Decision Tree model performs best in terms of MSE (0.86) and R^2 (0.12), while SVR follows closely with an MAE of 0.24 and R^2 of 0.1. Nu-SVR and Linear SVR achieve the lowest MAE (0.22), though Linear SVR has a slightly higher MSE (0.95). K-Neighbors Regression performs the worst, with an MSE of 1.08 and a negative R^2 (-0.11), indicating poor fit. Bayesian Ridge and SGDRegression show moderate performance, with MAPE values ranging from 0.65 to 0.96.

Regression models	MAE	MSE	MAPE	R2
SVR	0.24	0.88	0.63	0.1
Nu-SVR	0.22	0.93	0.66	0.05
Linear SVR	0.22	0.95	0.61	0.03
K-Neighbors Regression	0.34	1.08	1.25	-0.11
Decision tree regression	0.22	0.86	0.79	0.12
Bayesian Ridge	0.29	0.93	0.96	0.05
SGDRegression	0.24	0.96	0.65	0.017

Table 9. Regression results for the AI DB and without dimension reduction.

In Table 10, where PCA-based dimension reduction is applied, results vary. Nu-SVR improves slightly, achieving the lowest MAE (0.21) and MAPE (0.55), while its MSE remains stable (0.89). However, Linear SVR deteriorates significantly, with MSE increasing to 0.97 and R^2 dropping to -

0.004. Decision Tree regression, which performed well without PCA, now shows a higher MSE (1.02) and a negative R^2 (-0.04). K-Neighbors Regression remains poor, with nearly identical metrics as in Table 7. Bayesian Ridge is largely unaffected by PCA, maintaining similar MAE, MSE, and R^2 values.

Regression models	MAE	MSE	MAPE	R2
SVR	0.24	0.89	0.63	0.08
Nu-SVR	0.21	0.89	0.55	0.08
Linear SVR	0.25	0.97	0.94	-0.004
K-Neighbors Regression	0.34	1.08	1.28	-0.11
Decision tree regression	0.23	1.02	0.73	-0.04
Bayesian Ridge	0.29	0.93	0.96	0.04
SGDRegression	0.24	0.97	0.59	-0.0004

Table 10. Regression results for the AI DB and with dimension reduction based on PCA.

Tables 11 to 15 make a brief comparison of different proposed methods based on results of four performance metrics, including MAE, MSE, MAPE, and R2. In these tables, DR stands for dimension reduction. In our implementations, APJ DB has been evaluated with DR and without DR; AI DB follows the same rule as APJ DB with and without DR but both JI DB and PI BI have been used without DR.

Regression models	MAE	MSE	MAPE	R2
SVR	0.26	0.99	0.71	-0.023
Nu-SVR	0.24	1.02	0.73	-0.05
Linear SVR	0.26	0.99	0.73	-0.02
K-Neighbors Regression	0.26	0.92	1.06	0.05
Decision tree regression	0.23	1.02	0.71	-0.05
Bayesian Ridge	0.31	0.97	0.94	-0.001
SGDRegression	0.32	0.97	0.98	-0.001

Table 8. Regression results for the JI DB and without dimension reduction.

Regression models	MAE	MSE	MAPE	R2
SVR	0.11	0.21	0.29	0.78
Nu-SVR	0.09	0.16	0.24	0.84
Linear SVR	0.20	0.54	0.60	0.44
K-Neighbors Regression	0.17	0.42	0.51	0.56
Decision tree regression	0.16	0.46	0.48	0.52
Bayesian Ridge	0.19	0.55	0.6	0.42
SGDRegression	0.19	0.56	0.58	0.42

Table 9. Regression results for the PI DB and without dimension reduction.

Method	Min (MAE)	Algorithm
AJP DB-without DR	0.1558	Nu-SVR
AJP DB-with DR	0.1835	KNN
AI DB-without DR	0.2150	LSVR
AI DB-with DR	0.2088	Nu-SVR
JI DB-without DR	0.2339	Decision Tree
PI DB-without DR	0.0930	Nu-SVR

Table 10. Comparing all the methods based on MAE.

Method	Min (MSE)	Algorithm
AJP DB-without DR	0.5020	SVR
AJP DB-with DR	0.5300	KNN
AI DB-without DR	0.8570	Decision Tree
AI DB-with DR	0.8904	Nu-SVR
JI DB-without DR	0.9200	KNN
PI DB-without DR	0.1587	Nu-SVR

Table 11. Comparing all the methods based on MSE.

Method	Min (MAPE)	Algorithm
AJP DB-without DR	0.3618	Nu-SVR
AJP DB-with DR	0.5300	Nu-SVR
AI DB-without DR	0.6114	LSVR
AI DB-with DR	0.5517	Nu-SVR
JI DB-without DR	0.7158	Decision Tree
PI DB-without DR	0.2437	Nu-SVR

Table 12. Comparing all the methods based on MAPE.

Moreover, in Tables 16 and 17, we summarized all information in tables to find out which algorithm leads us to the best result regarding the methods. It can be clearly seen that the dominant algorithm is Nu-SVR, which outperformed all the other ones, and in terms of methods, features included in PI DB led us to the least error rates and the most R2.

Method	Max (R2)	Algorithm
AJP DB-without DR	0.4828	SVR
AJP DB-with DR	0.4545	KNN
AI DB-without DR	0.1173	Decision Tree
AI DB-with DR	0.0838	Nu-SVR
JI DB-without DR	0.0533	KNN
PI DB-without DR	0.8366	Nu-SVR

Table 13. Comparing all the methods based on R2.

Performance Metric	Method	Algorithm
MAE	PI DB-without DR	Nu-SVR
MSE	PI DB-without DR	Nu-SVR
MAPE	PI DB-without DR	Nu-SVR
R2	PI DB-without DR	Nu-SVR

Table 14. Comparing best-achieved results among all methods along with different performance metrics.

Furthermore, Figure 5 and Figure 6 draw a wide comparison among all proposed algorithms and our four introduced datasets in terms of MAE and R2, respectively.

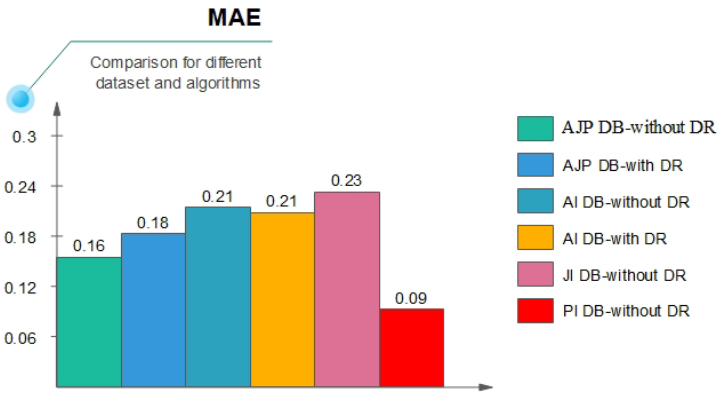

Figure 5. A comparison among all proposed algorithms and datasets based on MAE.

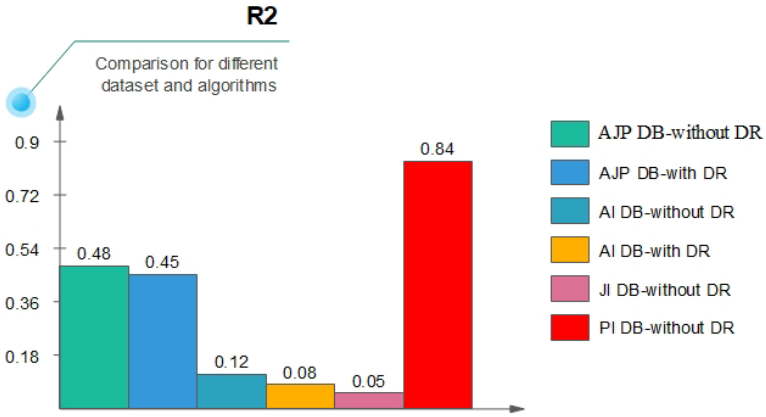

Figure 6. A comparison among all proposed algorithms and datasets based on R2.

Ultimately, we called our best model found on PI DB-without DR & Nu-SVR as PI-CCP. According to Tables 18, 19, and 20, the simulation results achieved based on the proposed PI-CCP modle are compared with the available studies presented by (Abrishami et al., 2019; Gao et al., 2024; Li et al., 2015; Li et al., 2019). Table 18 shows that PI-CCP performs exceptionally well at minimizing errors, as evidenced by the remarkably low MAE. It is noteworthy that it performs better than esteemed models like (Li et al., 2019), demonstrating its effectiveness in generating precise forecasts. The significant difference with (Gao et al., 2024) emphasizes PI-CCP's superiority even more.

Model	MAE
Proposed (PI-CCP)	0.0930
NIPS (S. Li et al., 2019)	0.1349
ICLR (S. Li et al., 2019)	0.1866
(Gao et al., 2024)	7.3000

Table 158. Achievement comparison based on MAE.

The R2 values demonstrate how well PI-CCP captures the variation in citation counts. surpassing both CCP and T-CCP (Li et al., 2015), and getting near NNCP (Abrishami & Aliakbary, 2019), PI-CCP demonstrates its validity as a predictor, underscoring its usefulness in figuring out the influence of research papers (see Table 19). Besides, Table 20 shows that PI-CCP maintains remarkable precision in citation count predictions, while NNCP (Abrishami & Aliakbary, 2019) displays a reduced MSE. This ensures a balance between accuracy and reliability. As a result, PI-CCP is

constantly positioned as a state-of-the-art and highly accurate model for citation count prediction by the thorough examination across MAE, R2, and MSE.

Model	R2
Proposed (PI-CCP)	0.83
CCP (C.-T. Li et al., 2015)	0.53
T-CCP (C.-T. Li et al., 2015)	0.68
NNCP (Abrishami & Aliakbary, 2019)	0.79

Table 169. Achievement comparison based on R2.

Model	MSE
Proposed (PI-CCP)	0.158
NNCP (Abrishami & Aliakbary, 2019)	0.034

Table 20. Achievement comparison based on MSE.

Conclusion and future work

Citation count serves as crucial metrics in evaluating the impact of scientific articles and researchers, playing a pivotal role in scholarly and academic endeavors. The purpose of our research was to implement a high-accuracy citation count prediction (CCP) model based on easy access and available public data. As a first step, we created four datasets (AI DB, JI DB, and PI DB) containing twenty three proposed attributes. The data were collected from about 2000 GSPs in the fields of computer science and electrical engineering. The obtained results of the proposed model, so called PI-CCP, allows us to conclude that the features suggested in the Paper Information Dataset (PI DB) including T_W^{Paper},$C_m^{Publication-year}$,$C_m^{Last-year}$,N^A, and 0AA are the most crucial variables in predicting the number of citations for scientific works. Besides, the most effective regression technique for forecasting citation count was determined to be the Nu-SVR algorithm. Achieving records such as 0.0930 for MAE and 0.8366 for R2 confirm that among all discussed algorithms, Nu-SVR is capable of managing the intricate, non-linear correlations between the input features and citation counts. Additionally, we examined APJ DB and AI DB in two different configurations: one with PCA/t-SNE-based dimension reduction and the other without them. The results showed that neither PCA nor t-SNE can consistently produce a lower error. Comparative analyses against existing models in the literature affirm the significant advancements achieved by our proposed algorithm, notably outperforming others. Our research not only presents the robust PI-CCP model but also contributes methodologically by offering insights into the selection of parameters, dataset creation, and algorithmic choices. The identification of crucial variables and the superior performance of Nu-SVR underscore the novelty and significance of our work in the domain of citation count prediction.

Even though we introduced twenty three novel features, there may still be other potentially relevant variables that could not be missed. Factors such as historical citation trends (time series), collaboration networks, social media presence, the impact of conference versus journal publications, institutional and funding factors, etc. may also play significant roles in citation counts but were not included in our analysis. Besides, our study focuses specifically on computer science and electrical engineering. The findings may not be directly applicable to other fields, as citation behaviors can vary significantly across disciplines. Future research could explore the applicability of our model in different academic domains.

Disclosure statement: The authors report there are no competing interests to declare.

Funding statement: There is no funding resource for this study.

Conflict of interest: There is no conflict of interest to declare.

About the authors

Mahdi Bahaghighat is an Associate Professor in the Department of Computer Engineering at Imam Khomeini International University, Qazvin, Iran. His primary research interests include Artificial Intelligence, Computer Vision, Natural Language Processing, and applications of AI in Finance. Dr. Bahaghighat leads the Artificial Intelligence in Science and Technologies (AIST) laboratory. As the corresponding author, he can be reached via email at Bahaghighat@eng.ikiu.ac.ir.

Leila Akbari is a Research Assistant in the AIST Lab. She holds a M.Sc. in Electrical Engineering from Islamic Azad University, Qazvin, Iran. Her research interests include artificial intelligence, signal processing, and image processing.

Majid Ghasemi is an undergraduate student in Computer Engineering Department, Imam Khomeini International University, Qazvin, Iran. His research interests include artificial intelligence, and machine learning.

Qin Xin is a Full Professor of Computer Science at the University of the Faroe Islands. He earned his Ph.D. from the University of Liverpool in 2004 and has held research positions at renowned institutions, including Simula Research Laboratory and UCLouvain. With over 200 peer-reviewed publications, his research focuses on algorithms for wireless networks, cryptography, and combinatorial optimization.

References

Abramo, G., D'angelo, C. A., & Di Costa, F. (2023). Correlating article citedness and journal impact: An empirical investigation by field on a large-scale dataset. *Scientometrics*, *128*(3), 1877–1894. https://doi.org/10.1007/s11192-022-04622-0

Abrishami, A., & Aliakbary, S. (2019). Predicting citation counts based on deep neural network learning techniques. *Journal of Informetrics*, *13*(2), 485–499. https://doi.org/10.1016/j.joi.2019.02.011

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *Sage Open*, *9*(1). https://doi.org/10.1177/215824401982957

Almas, K., Ur Rehman, S., Al-Harbi, F., Qadir Khan, S., Ahmed Farooqi, F., Smith, S., & Ahmad, S. (2021). Significance of variable contributing factors on impact factor of Clarivate analytics dental journals. *Serials Review*, *47*(3–4), 201–214. https://doi.org/10.1080/00987913.2021.2018225

Amin, M., & Mabe, M. A. (2003). Impact factors: Use and abuse. *Medicina (Buenos Aires)*, *63*(4), 347–354. https://medicinabuenosaires.com/revistas/vol63-03/4/Impact%20factors-use%20and%20abuse.pdf (Archived at https://web.archive.org/web/20250325222750/http://www.medicinabuenosaires.com/revistas/vol63-03/4/Impact%20factors-use%20and%20abuse.pdf)

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, 1(1), 377–386. https://doi.org/10.1162/qss_a_00019

Bahaghighat, Mahdi; Jahani rad, P. (2024). *AoI2WoS: Mapping area of interest in Google Scholar profile to Web Of Science (WoS) scientific fields categories*. Mendeley Data. http://doi.org/10.17632/nr7zfdjm7f.1

Rad, P. J., & Bahaghighat, M. (2024). *Hierarchical text classification for web of science scientific fields*. Facta Universitatis, Series: Electronics and Energetics, 37(4), 703-732.

https://doi.org/10.2298/FUEE2404703J

Bai, X., Zhang, F., Liu, J., Wang, X., & Xia, F. (2025). Revolutionizing scholarly impact: Advanced evaluations, predictive models, and future directions. *SpringerX Bai, F Zhang, J Liu, X Wang, F XiaArtificial Intelligence Review*, 2025•*Springer*, 58(10). https://doi.org/10.1007/S10462-025-11315-6

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10), 203–224. https://static.aminer.org/pdf/PDF/000/337/560/uncertainty_support_vector_method_for_ordinal_regression.pdf (Archived at https://web.archive.org/web/20200709085151/https://static.aminer.org/pdf/PDF/000/337/560/uncertainty_support_vector_method_for_ordinal_regression.pdf)

Belikov, A. V, & Belikov, V. V. (2015). A citation-based, author-and age-normalized, logarithmic index for evaluation of individual researchers independently of publication counts. *F1000Research*, 4. https://doi.org/10.12688/f1000research.7070.2

Bhatt, D., Aggarwal, P., Bhattacharya, P., & Devabhaktuni, V. (2012). An enhanced mems error modeling approach based on nu-support vector regression. *Sensors*, 12(7), 9448–9466. https://doi.org/10.3390/s120709448

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer Google Scholar, 2, 1122–1128.

Blümel, C., & Schniedermann, A. (2020). Studying review articles in scientometrics and beyond: A research agenda. *Scientometrics*, 124(1), 711–728. https://doi.org/10.1007/s11192-020-03431-7

Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), 45–80. https://doi.org/10.1108/00220410810844150

Bornmann, L., & Daniel, H. (2009). The state of h index research: Is the h index the ideal way to measure research performance? *EMBO Reports*, 10(1), 2–6. https://doi.org/10.1038/embor.2008.233

Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics*, 69, 169–173. https://doi.org/10.1007/s11192-006-0147-4

Broadus, R. N. (1987). Toward a definition of “bibliometrics.” *Scientometrics*, 12(5–6), 373–379. https://doi.org/10.1007/BF02016680

Butler, L., & Visser, M. S. (2006). Extending citation analysis to non-source items. *Scientometrics*, 66(2), 327–343.

Cameron, A. C., & Windmeijer, F. A. G. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2), 329–342. https://doi.org/10.1016/S0304-4076(96)01818-0

Cao, X., Chen, Y., & Liu, K. J. R. (2016). A data analytic approach to quantifying scientific impact. *Journal of Informetrics*, 10(2), 471–484. https://doi.org/10.1016/j.joi.2016.02.006

Chen, J. T., Lee, C., & Chen, L. Y. (2024). *Statistical prediction and machine learning*. CRC Press.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Cunningham, P., & Delany, S. J. (2021). K-nearest neighbour classifiers-a tutorial. *ACM Computing Surveys (CSUR)*, 54(6), 1–25. https://doi.org/10.1145/3459665

Durieux, V., & Gevenois, P. A. (2010). Bibliometric indicators: Quality measurements of scientific publication. *Radiology*, 255(2), 342–351. https://doi.org/10.1148/radiol.09090626

Enduri, M. K., Sankar, V. U., & Hajarathaiah, K. (2022). Empirical study on citation count prediction of research articles. *Journal of Scientometric Research*, 11(2), 155–163. https://doi.org/10.5530/jscires.11.2.17

Fassin, Y. (2020). The HF-rating as a universal complement to the h-index. *Scientometrics*, 125(2), 965–990. https://doi.org/10.1007/s11192-020-03611-5

Fox, J. (2015). *Applied regression analysis and generalized linear models*. SAGE Publications, Inc.

Furman, J. L., & Teodoridis, F. (2020). Automation, research technology, and researchers' trajectories: Evidence from computer science and electrical engineering. *Organization Science*, 31(2), 330–354. https://doi.org/10.1287/orsc.2019.1308

Gao, T., Liu, J., Pan, R., & Wang, H. (2024). Citation counts prediction of statistical publications based on multi-layer academic networks via neural network model. *Expert Systems with Applications*, 238, 121634. https://doi.org/10.1016/j.eswa.2023.121634

Garfield, E. (2006). The history and meaning of the journal impact factor. *Jama*, 295(1), 90–93. https://doi.org/10.1001/jama.295.1.90

González-Betancor, S. M., & Dorta-González, P. (2017). An indicator of the impact of journals based on the percentage of their highly cited publications. *Online Information Review*, 41(3), 398–411. https://doi.org/10.1108/OIR-01-2016-0008

Groos, O. V, & Pritchard, A. (1969). Documentation notes. *Journal of Documentation*, 25(4), 344–349. https://doi.org/10.1108/eb026482

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications. *Journal of Big Data*, 11(1), 113. https://doi.org/10.1186/s40537-024-00973-y

He, G., Gu, S., Xue, Z., Duan, Y., & Zhu, X. (2025). Sequential citation counts prediction enhanced by dynamic contents. *Journal of Informetrics*, 19(2), 101645. https://doi.org/10.1016/j.joi.2025.101645

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, 13(4), 18–28. https://doi.org/10.1109/5254.708428

Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, 85(3), 741–754. https://doi.org/10.1007/s11192-010-0193-9

Hutchins, B. I., Yuan, X., Anderson, J. M., & Santangelo, G. M. (2016). Relative citation ratio (RCR): A new metric that uses citation rates to measure influence at the article level. *PLoS Biology*, 14(9). https://doi.org/10.1371/journal.pbio.1002541

Jung, S., Dagobert, T., Morel, J.-M., & Facciolo, G. (2024). A review of t-SNE. *Image Processing On Line*, 14, 250–270. https://doi.org/10.5201/ipol.2024.528

Khokhlov, A. N. (2020). How scientometrics became the most important science for researchers of all specialties. *Moscow University Biological Sciences Bulletin*, 75, 159–163. https://doi.org/10.3103/s0096392520040057

Khurana, P., & Sharma, K. (2022). Impact of h-index on author's rankings: An improvement to the h-index for lower-ranked authors. *Scientometrics*, 127(8), 4483–4498. https://doi.org/10.1007/s11192-022-04464-w

Kosyakov, D., & Pislyakov, V. (2024). "I'd like to publish in Q1, but there's no Q1 to be found": Study of journal quartile distributions across subject categories and topics. *Journal of Informetrics*, 18(1), 101494. https://doi.org/10.1016/j.joi.2024.101494

Li, C.-T., Lin, Y.-J., Yan, R., & Yeh, M.-Y. (2015). Trend-based citation count prediction for research articles. *Advances in Knowledge Discovery and Data Mining: 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings, Part I* 19, 659–671. http://dx.doi.org/10.1007/978-3-319-18038-0_51

Li, S., Zhao, W. X., Yin, E. J., & Wen, J.-R. (2019). A neural citation count prediction model based on peer review text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4914–4924. https://doi.org/10.18653/v1/D19-1497

Lundberg, J. (2006). *Bibliometrics as a research assessment tool: impact beyond the impact factor*. Karolinska Institutet (Sweden).

Moed, H. F. (2006). *Citation analysis in research evaluation* (Vol. 9). Springer Science & Business Media.

Moussa, S. (2023). A bibliometric investigation of the journals that were repeatedly suppressed from Clarivate's journal citation reports. *Accountability in Research*, 30(8), 592–612. https://doi.org/10.1080/08989621.2022.2071154

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116(12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116%3C2417:SSBOTM%3E2.0.CO;2

Nguyen, B. T., & Nguyen, T. T. (2025). Forecasting scientific impact: A model for predicting citation counts. *Statistics, Optimization & Information Computing*, 13(6), 2601–2615. https://doi.org/10.19139/soic-2310-5070-2524

Okagbue, H. I., Akhmetshin, E. M., & Teixeira da Silva, J. A. (2021). Distinct clusters of CiteScore and percentiles in top 1000 journals in Scopus. *COLLNET Journal of Scientometrics and Information Management*, 15(1), 133–143. https://doi.org/10.1080/09737766.2021.1934604

Okagbue, H. I., Bishop, S. A., Adamu, P. I., Opanuga, A. A., & Obasi, E. C. M. (2020). Analysis of percentiles of computer science, theory and methods journals: CiteScore versus impact factor. *DESIDOC Journal of Library & Information Technology*, 40(1), 359–365. http://dx.doi.org/10.14429/djlit.40.1.14866

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Pobiedina, N., & Ichise, R. (2016). Citation count prediction as a link prediction problem. *Applied Intelligence*, 44, 252–268. https://doi.org/10.1007/s10489-015-0657-y

Rostami, M., Bahaghighat, M., & Zanjireh, M. M. (2021). Bitcoin daily close price prediction using optimized grid search method. *Acta Universitatis Sapientiae, Informatica*, 13(2), 265–287. https://doi.org/10.2478/ausi-2021-0012

Sabry, F. (2023). *K Nearest Neighbor algorithm: Fundamentals and applications* (Vol. 28). One Billion Knowledgeable.

Skrodzki, M., van Geffen, H., Chaves-de-Plaza, N. F., Höllt, T., Eisemann, E., & Hildebrandt, K. (2024). Accelerating hyperbolic t-SNE. *IEEE Transactions on Visualization and Computer Graphics*, 30(7), 4403–4415. https://doi.org/10.1109/TVCG.2024.3364841

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Sohrabi, B., & Iraj, H. (2017). The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts. *Scientometrics*, 110, 243–251. https://doi.org/10.1007/s11192-016-2161-5

Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing*, 251, 26–34. https://doi.org/10.1016/j.neucom.2017.04.018

Teixeira da Silva, J. A. (2020). CiteScore: Advances, evolution, applications, and limitations. *Publishing Research Quarterly*, 36(3), 459–468. https://doi.org/10.1007/s12109-020-09736-y

Torres-Salinas, D., Valderrama-Baca, P., & Arroyo-Machado, W. (2022). Is there a need for a new journal metric? Correlations between JCR Impact Factor metrics and the Journal Citation Indicator—JCI. *Journal of Informetrics*, 16(3), 101315. https://doi.org/10.1016/j.joi.2022.101315

Wallisch, C., Bach, P., Hafermann, L., Klein, N., Sauerbrei, W., Steyerberg, E. W., Heinze, G., Rauch, G., & Initiative, T. G. 2 of the S. (2022). Review of guidance papers on regression modeling in statistical series of medical journals. *PloS One*, 17(1). https://doi.org/10.1371/journal.pone.0262918

Wang, B., Wu, F., & Shi, L. (2023). AGSTA-NET: Adaptive graph spatiotemporal attention network for citation count prediction. *Scientometrics*, 128(1), 511–541. https://doi.org/10.1007/s11192-022-04541-0

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79–82. https://doi.org/10.3354/cr030079

Yao, Z., & Ruzzo, W. L. (2006). A regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC Bioinformatics*, 7, 1–11. https://doi.org/10.1186/1471-2105-7-S1-S11

Yu, T., Yu, G., Li, P.-Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics*, 101, 1233–1252. https://doi.org/10.1007/s11192-014-1279-6

Zafar, L., Masood, N., Hadi, F., & Ahmed, S. (2024). Citation count prediction of scholarly articles. *Journal of Computing & Biomedical Informatics*, 6(2). https://doi.org/10.56979/602/2024

Zhang, Z., Yu, C., Wang, J., & An, L. (2025). A temporal evolution and fine-grained information aggregation model for citation count prediction. *Scientometrics*, 130(4), 2069–2091. https://doi.org/10.1007/s11192-025-05294-2

Zhu, J., Zhou, J., Pan, J., Gu, F., & Guo, J. (2025). Ranking influential non-content factors on scientific papers' citation impact: A multidomain comparative analysis. *Big Data and Cognitive Computing*, 9(2), 30. https://doi.org/10.3390/bdcc9020030

Copyright