Future Internet • *Open Access* • Volume 14, Issue 8 • August 2022 • Article number 229

Document type
Article • *Gold Open Access*

Source type
Journal

ISSN
19995903

DOI
10.3390/fi14080229

View more

CCrFS: Combine Correlation Features Selection for Detecting Phishing Websites Using Machine Learning

Moedjahedy, Jimmy[a] ; Setyanto, Arief[b] ;
Alarfaj, Fawaz Khaled[c] ; Alreshoodi, Mohammed[d]
Save all to author list

[a] Computer Science Department, Universitas Klabat, Minahasa Utara, 95371, Indonesia
[b] Magister of Informatics Engineering, Universitas AMIKOM Yogyakarta, Yogyakarta, 55281, Indonesia
[c] Department of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh, 11564, Saudi Arabia
[d] Unit of Scientific Research, Applied College, Qassim University, Buraydah, 52362, Saudi Arabia

Abstract

Internet users are continually exposed to phishing as cybercrime in the 21st century. The objective of phishing is to obtain sensitive information by deceiving a target and using the information for financial gain. The information may include a login detail, password, date of birth, credit card number, bank account number, and family-related information. To acquire these details, users will be directed to fill out the information on false websites based on information from emails, adverts, text messages, or website pop-ups. Examining the website's URL address is one method for avoiding this type of deception. Identifying the features of a phishing website URL takes specialized knowledge and investigation. Machine learning is one method that uses existing data to teach machines to distinguish between legal and phishing website URLs. In this work, we proposed a method that combines correlation and recursive feature elimination to determine which URL characteristics are useful for identifying phishing websites by gradually decreasing the number of features while maintaining accuracy value. In this paper, we use two datasets that contain 48 and 87 features. The first scenario combines power predictive score correlation and recursive feature elimination; the second scenario is the maximal information coefficient correlation and recursive feature elimination. The third scenario combines spearman correlation and recursive feature elimination. All three scenarios from the combined findings of the proposed methodologies achieve a high level of accuracy even with the smallest feature subset. For dataset 1, the accuracy value for the 10 features result is 97.06%, and for dataset 2 the accuracy value is 95.88% for 10 features. © 2022 by the authors.

Author keywords

correlation; feature elimination; feature selection; machine learning; phishing detection

Indexed keywords

Future Internet

COUNTRY	SUBJECT AREA AND CATEGORY	PUBLISHER	H-INDEX
Switzerland Universities and research institutions in Switzerland	Computer Science Computer Networks and Communications	Multidisciplinary Digital Publishing Institute (MDPI)	38

PUBLICATION TYPE	ISSN	COVERAGE	INFORMATION
Journals	19995903	2009-2021	Homepage How to publish in this journal dino.giuli@unifi.it

SCOPE

Macro-Area 1: Smart System Technologies and Architecture. Included topics are: • advanced communications network infrastructures • evolution of internet basic services • internet of things • netted peripheral sensors • industrial internet • centralized and distributed data centers • embedded computing • cloud computing • software defined network functions and network virtualization • cloud-let and fog-computing • big data, open data and analytical tools • cyber-physical systems • network and distributed operating systems • web services • semantic structures and related software tools • artificial and augmented intelligence • augmented reality • system interoperability and flexible service composition • smart mission-critical system architectures • smart terminals and applications • pro-sumer tools for application design and development • cyber security compliance • privacy compliance • reliability compliance • dependability compliance • accountability compliance • trust compliance • technical quality of basic services Macro-Area 2: Smart Systems and Applications. Included topics are: • smart mobility and transportation systems • smart utility systems • smart energy systems • smart living places • smart public government systems • smart health-care systems • smart systems for public security and safety • smart social assistance systems[...] Macro-Area 3: Net-Living Human Factors and Quality of Life enhancement [...]

Join the conversation about this journal

Quartiles

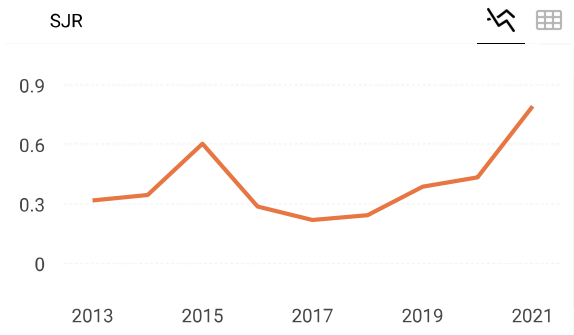
FIND SIMILAR JOURNALS

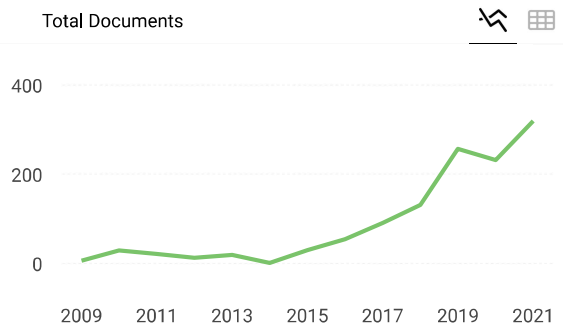
1
Mobile Information Systems
EGY
74%
similarity

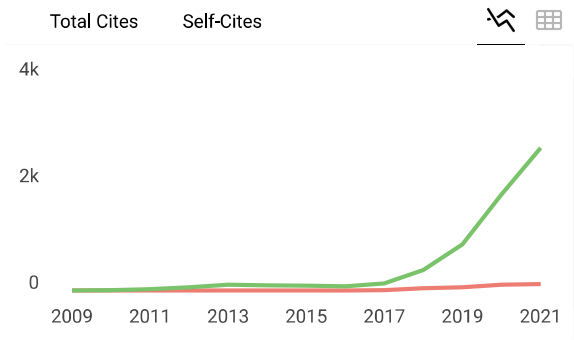
2
Mobile Networks and Applications
NLD
64%
similarity

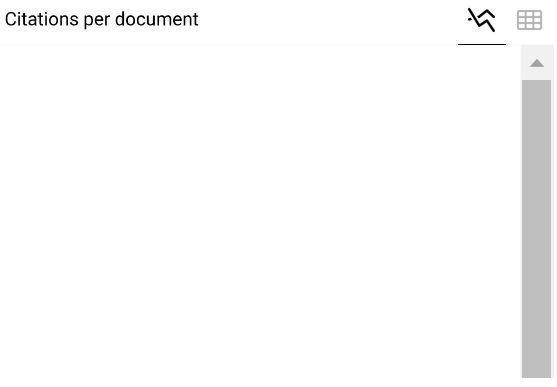
3
Annales des Telecommunications/Annals
FRA
62%
similarity

4
Pervasive and M Computing
NLD
62
simila

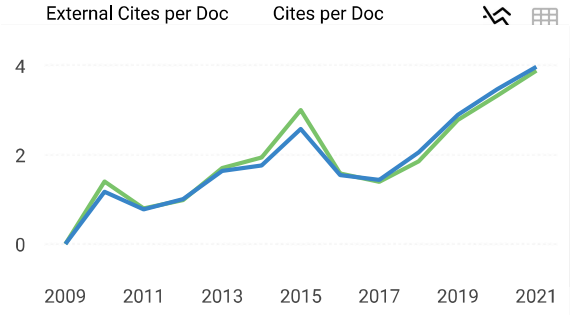




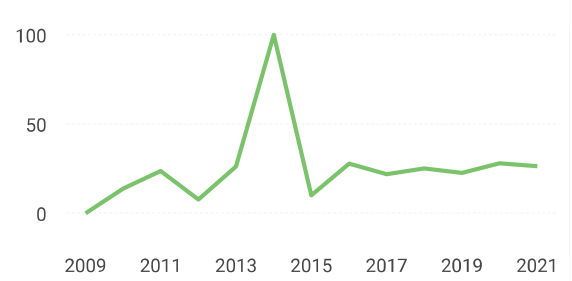




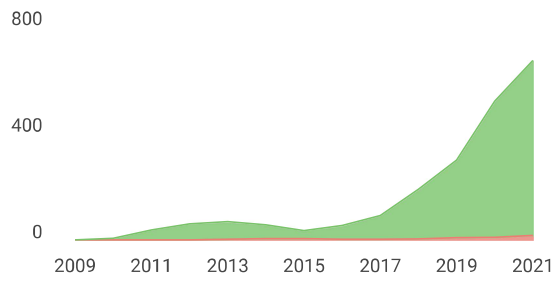
External Cites per Doc Cites per Doc



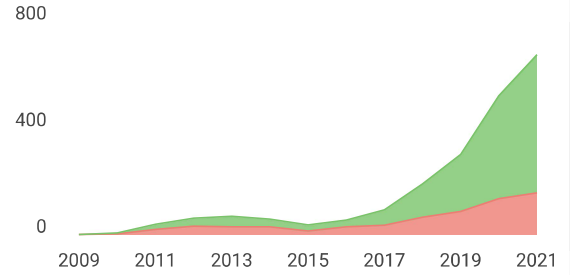
% International Collaboration



Citable documents Non-citable documents



Cited documents Uncited documents



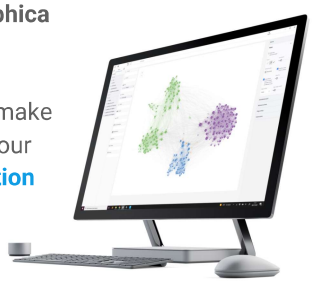
← Show this widget in your own website

Just copy the code below and paste within your html code:

```
<a href="https://www.scimag
```

SCImago Graphica

Explore, visually communicate and make sense of data with our **new data visualization tool**.



Metrics based on Scopus® data as of April 2022

S **saad** 3 years ago

hi

reply

2021 Journal Performance Data for: Future Internet

Open Access since 2009

ISSN
N/A

EISSN
1999-5903

JCR ABBREVIATION
FUTURE INTERNET

ISO ABBREVIATION
Future Internet

Journal Information

EDITION
Emerging Sources Citation Index (ESCI)

CATEGORY
COMPUTER SCIENCE, INFORMATION SYSTEMS - ESCI

LANGUAGES
English

REGION
SWITZERLAND

1ST ELECTRONIC JCR YEAR
2020

Publisher Information

PUBLISHER
MDPI

ADDRESS
ST ALBAN-ANLAGE 66, CH-4052 BASEL, SWITZERLAND

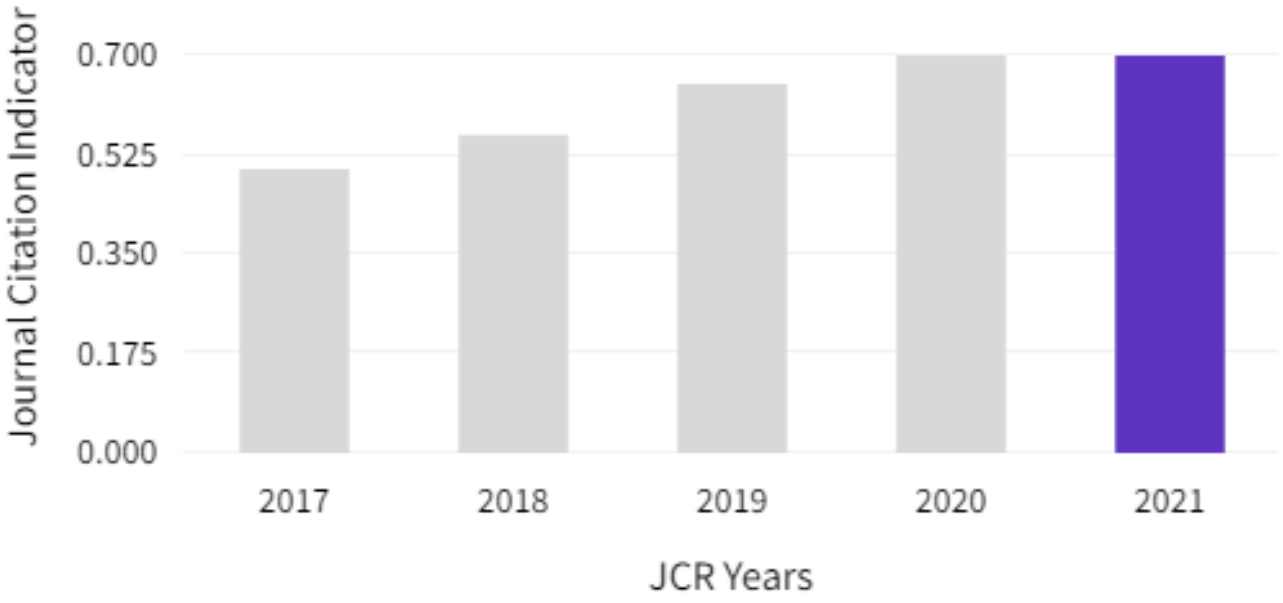
PUBLICATION FREQUENCY
12 issues/year

Journal's Performance

Journal Citation Indicator (JCI)

0.70

The Journal Citation Indicator (JCI) is the average Category Normalized Citation Impact (CNCI) of citable items (articles & reviews) published by a journal over a recent three year period. The average JCI in a category is 1. Journals with a JCI of 1.5 have 50% more citation impact than the average in that category. It may be used alongside other metrics to help you evaluate journals.



Total Citations

2,736

The total number of times that a journal has been cited by all journals included in the database in the JCR year. Citations to journals listed in JCR are compiled annually from the JCR years combined database, regardless of which JCR edition lists the journal.

Open Access (OA)

The data included in this tile summarizes the items published in the journal in the JCR data year and in the previous two years. For example, in the 2020 JCR data, released in June 2021, the Open Access (OA) data show the publication model (Gold OA or subscription) of materials published in 2018, 2019 and 2020, and citations in 2020 to these items. This three-year set of published items is used to provide descriptive analysis of the content and community of the journal.

Items

TOTAL CITABLE
794

% OF CITABLE OA
95.84%

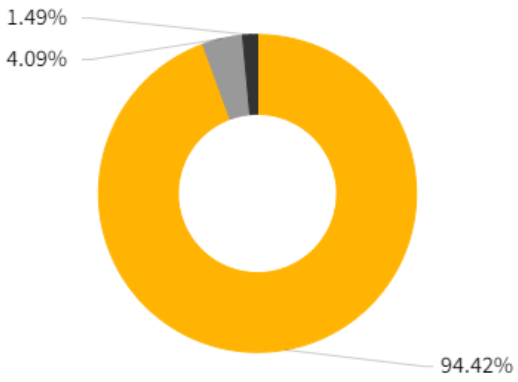
CITABLE

GOLD OPEN ACCESS
761 / 94.42%

SUBSCRIPTION OR BRONZE
33 / 4.09%

NON-CITABLE

OTHER (NON-CITABLE ITEMS)
12 / 1.49%



Citations*

TOTAL CITABLE
1,724

% OF CITABLE OA
96.52%

CITABLE

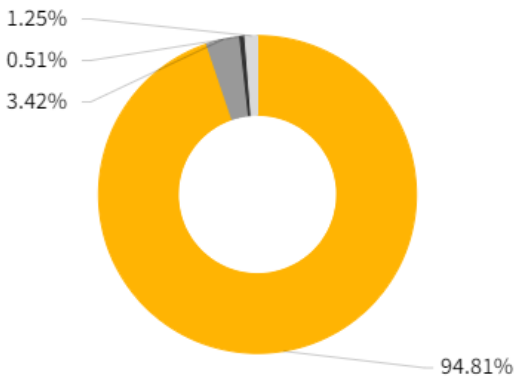
GOLD OPEN ACCESS
1,664 / 94.81%

SUBSCRIPTION OR BRONZE
60 / 3.42%

NON-CITABLE

OTHER (NON-CITABLE ITEMS)
9 / 0.51%

UNLINKED CITATIONS
22 / 1.25%



* Citations in 2021 to items published in (2019-2021)

Rank by Journal Citation Indicator (JCI)

Journals within a category are sorted in descending order by Journal Citation Indicator (JCI) resulting in the Category Ranking below. A separate rank is shown for each category in which the journal is listed in JCR. Data for the most recent year is presented at the top of the list, with other years shown in reverse chronological order.

Only journals which have a calculated JCI value are included in the JCI ranking. The total number of journals displayed in this ranking may be less than the category overall.

CATEGORY

COMPUTER SCIENCE, INFORMATION SYSTEMS

115/246

JCR YEAR	JCI RANK	QUARTILE	JCI PERCENTILE
2021	115/246	Q2	53.46
2020	96/223	Q2	57.17
2019	111/223	Q2	50.45
2018	128/220	Q3	42.05
2017	138/216	Q3	36.34

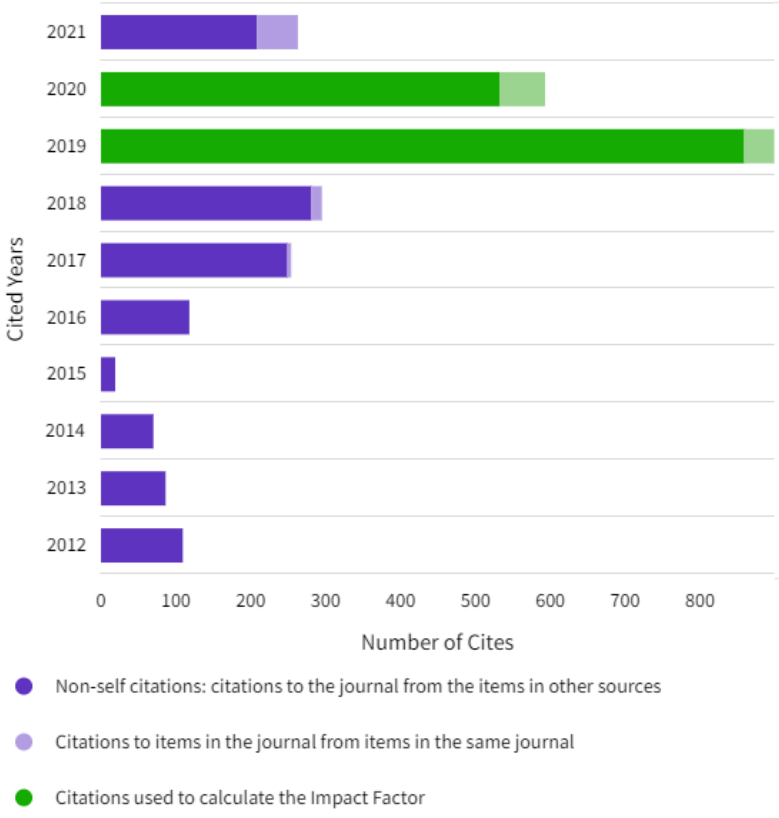
Citation network

Cited Half-life

2.6 years

The Cited Half-Life is the median age of the items in this journal that were cited in the JCR year. Half of a journal's cited items were published more recently than the cited half-life.

TOTAL NUMBER OF CITES	NON-SELF CITATIONS	SELF CITATIONS
2,736	2,558	178



	# OF CITES FROM 2021	CUMULATIVE %	# OF CITING SOURCES
	2,736 citations	**100.00%**	**894 sources**
2021	263 citations	9.61%	105 sources
2020	593 citations	31.28%	249 sources
2019	899 citations	64.14%	376 sources
2018	295 citations	74.92%	191 sources
2017	254 citations	84.20%	152 sources
2016	118 citations	88.51%	105 sources
2015	19 citations	89.20%	19 sources
2014	70 citations	91.76%	62 sources
2013	87 citations	94.94%	72 sources
2012	110 citations	98.96%	92 sources
Previous years:	28 citations		

Citing titles in all years

Future Internet

	SOURCE NAME	COUNT
	All Others	574
1	Future Internet	178
2	SENSORS	157
3	IEEE Access	156
4	Sustainability	99
5	Applied Sciences-Basel	86
6	Electronics	81
7	LECT NOTES COMPUT SC	25
8	ISPRS International Journal of Geo-Information	21
9	Computer Networks	20
10	International Journal of Advanced Computer Science and Applications	20
11	International Journal of Environmental Research and Public Health	19
12	WIRELESS PERSONAL COMMUNICATIONS	19
13	Information	18
14	PeerJ Computer Science	18
15	COMPUTERS & SECURITY	17
16	IEEE Internet of Things Journal	17
17	IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS	17
18	Security and Communication Networks	17
19	Education Sciences	16
20	Energies	16

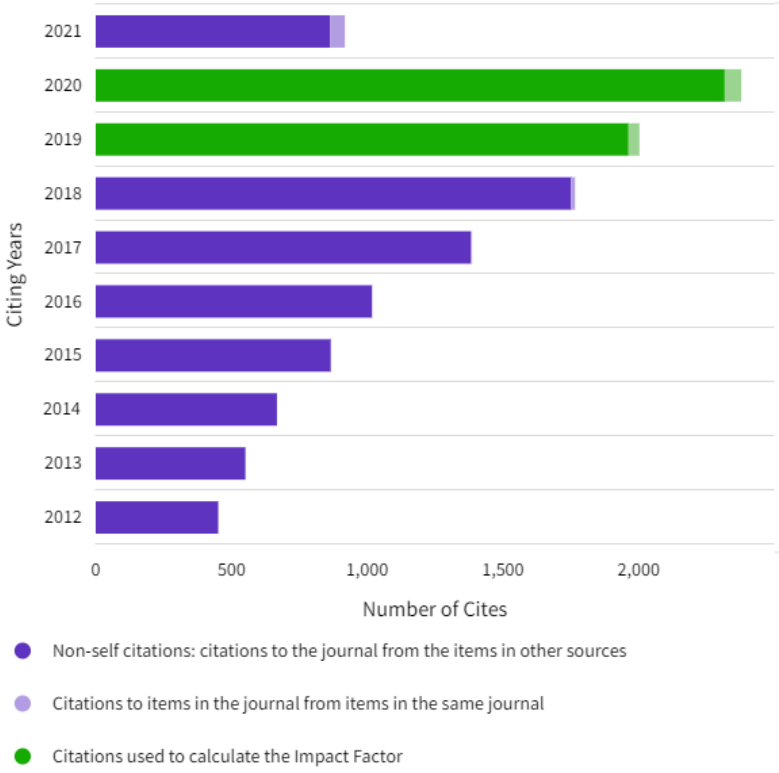
Showing 1 - 20 rows of 320 total (use export in the relevant section to download the full table)

Citing Half-life

4.7 years

The Citing Half-Life is the median age of items in other publications cited by this journal in the JCR year.

TOTAL NUMBER OF CITES	NON-SELF CITATIONS	SELF CITATIONS
16,076	15,898	178



	# OF CITES FROM 2021	CUMULATIVE %	# OF CITED SOURCES
	16,076 citations	**100.00%**	**8,453 sources**
2021	917 citations	5.70%	579 sources
2020	2,376 citations	20.48%	1,379 sources
2019	2,001 citations	32.93%	1,219 sources
2018	1,764 citations	43.90%	1,159 sources
2017	1,385 citations	52.52%	958 sources
2016	1,017 citations	58.85%	743 sources
2015	865 citations	64.23%	642 sources
2014	667 citations	68.38%	512 sources
2013	552 citations	71.81%	455 sources
2012	452 citations	74.62%	381 sources
Previous years:	4,080 citations		

Cited titles in all years

Future Internet

	SOURCE NAME	COUNT
	All Others	6,851
1	IEEE Access	385
2	Future Internet	178
3	LECT NOTES COMPUT SC	152
4	SENSORS	142
5	PROC CVPR IEEE	136
6	IEEE Internet of Things Journal	116
7	COMPUTERS IN HUMAN BEHAVIOR	102
8	Sustainability	87
9	IEEE Communications Surveys and Tutorials	74
10	Applied Sciences-Basel	71
11	IEEE COMMUNICATIONS MAGAZINE	68
12	EXPERT SYSTEMS WITH APPLICATIONS	65
13	Future Generation Computer Systems-The International Journal of eScience	61
14	IEEE I CONF COMP VIS	60
15	IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY	55
16	COMPUTERS & EDUCATION	50
17	Computer Networks	49
18	PLoS One	49
19	PROCEDIA COMPUT SCI	43
20	SCIENCE	42

Showing 1 - 20 rows of 1602 total (use export in the relevant section to download the full table)

Content metrics

Source data

This tile shows the breakdown of document types published by the journal. Citable Items are Articles and Reviews. For the purposes of calculating JIF, a JCR year considers the publications of that journal in the two prior years.

312 total citable items

	ARTICLES	REVIEWS	COMBINED (C)	OTHER DOCUMENT TYPES (O)	PERCENTAGE
NUMBER IN JCR YEAR 2021 (A)	286	26	312	1	100%
NUMBER OF REFERENCES (B)	13,390	2,672	16,062	14	100%
RATIO (B/A)	46.8	102.8	51.5	14.0	

Contributions by Organizations

Organizations that have contributed the most papers to the journal in the most recent three-year period.

RANK	ORGANIZATION	COUNT
1	SHANGHAI UNIVERSITY	27
2	UNIVERSITY OF FLORENCE	21
3	LEAGUE OF EUROPEAN RESEARCH UNIVERSITIES - LERU	15
4	CHINESE ACADEMY OF SCIENCES	13
5	TECNOLOGICO DE MONTERREY	11
6	NATIONAL TECHNICAL UNIVERSITY OF ATHENS	9
7	EGYPTIAN KNOWLEDGE BANK (EKB)	8
-	JILIN UNIVERSITY	8

Showing 1 - 8 rows of 952 total (use export in the relevant section to download the full table)

Contributions by country/region

Countries or Regions that have contributed the most papers to the journal in the most recent three-year period.

RANK	COUNTRY/REGION	COUNT
1	CHINA MAINLAND	185
2	Italy	119
3	USA	62
4	Greece	52
5	England	49
6	GERMANY (FED REP GER)	42
7	Spain	41
8	Russia	28

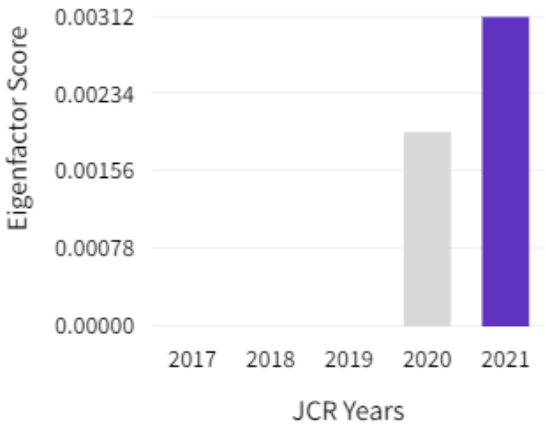
Showing 1 - 8 rows of 84 total (use export in the relevant section to download the full table)

Additional metrics

Eigenfactor score

0.00312

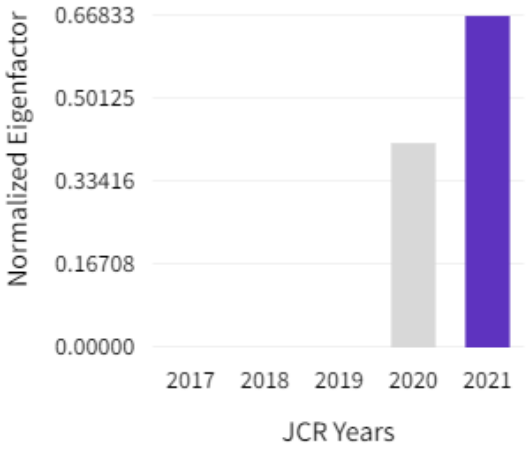
The Eigenfactor Score is a reflection of the density of the network of citations around the journal using 5 years of cited content as cited by the Current Year. It considers both the number of citations and the source of those citations, so that highly cited sources will influence the network more than less cited sources. The Eigenfactor calculation does not include journal self-citations.



Normalized Eigenfactor

0.66833

The Normalized Eigenfactor Score is the Eigenfactor score normalized, by rescaling the total number of journals in the JCR each year, so that the average journal has a score of 1. Journals can then be compared and influence measured by their score relative to 1.



Article influence score

0.440

The Article Influence Score normalizes the Eigenfactor Score according to the cumulative size of the cited journal across the prior five years. The mean Article Influence Score for each article is 1.00. A score greater than 1.00 indicates that each article in the journal has above-average influence.