Journal of Scientometric Research, 2023; 12(1):44-53.
https://www.jscires.org

Research Article

An Intelligent Prediction of the Next Highly Cited Paper Using Machine Learning

Galal M. Bin Makhashen*, Hamdi A. Al-Jamimi

Research Institute, King Fahd University of Petroleum and Minerals, Dhahran, SAUDI ARABIA.

ABSTRACT

Highly cited articles capture the attention of significant contributors in the research community as an opportunity to improve knowledge, source of ideas or solutions, and advance their research in general. Typically, these articles are authored by a large number of scientists with international collaboration. However, this could not be the only reason for an article to be highly cited, there might be several other characteristics for an article to be more attractive to researchers and readers. In other words, there are a few other characteristics that help articles/papers to be more than others to appear in search engines or to grab readers' attention. In this study, we modeled several machine-learning methods with a set of articles, and journal characteristics including authors-count, title characteristics, abstract length, international collaboration, number of keywords, funding information, journal characteristics, etc. We extracted 20 characteristics and developed multiple machine-learning models to automate highly-cited papers recognition from regular papers. In experiments conducted with an ensemble machine learning algorithm, 97% recognition accuracy was achieved. Other algorithms including a deep learning method using LSTMs also achieved high recognition accuracy. Such high performances can be utilized for a promising HCP auto-detection system in the future.

Keywords: Artificial Intelligence, Machine Learning, Highly Cited Paper Indicators, Digital Libraries, Bibliometric Analysis.

Correspondence:
Galal M BinMakhashen
Research Institute, King Fahd University of Petroleum and Minerals, Dhahran-31261, SAUDI ARABIA.
Email: binmakhashen@kfupm.edu.sa
ORCID ID: 0000-0002-5111-9760

Received: 13-06-2022;
Revised: : 08-09-2022;
Accepted: 01-01-2023.

INTRODUCTION

Scientific articles that are heavily cited by researchers and stand out as highly cited papers (HCPs) can be attributed to several factors. Among the well-known reasons is the international collaboration. Moreover, scientific papers with multi-country authors are in general cited more than single-country papers.[1] In general, HCP articles are important to the research community as an indicator of trending science.[2]

In the past two decades, there has been a significant focus on the analysis of it may affect the research funding strategy that changes, as a result, can help research administrations decide on the next research focus.[2,3] The HCP is one of the indicators used to identify quality research publication in the context of scientific excellence. Moreover, the HCP has been extended to quantify research performance at the institutional, university, and national levels.[4]

According to a European Commission benchmarking study, there is a growing interest in using highly cited papers as a metric of "world-class" scientific research.[2] HCP has been applied as an indicator for comparing the research performance of the EU countries since 2001. However, the application of citation count as an indicator is still controversial on how the HCP metric is related to scientific excellence.

The citation can be divided into positive or negative. Let us suppose that the positive citation occurs when one study refers to a previously published study with an agreement. On the other hand, a negative citation is uncommon because it is usually a formal expression of disagreement with the previous publication's content.[5] Compared to "Regular" Cited Papers (RCP), highly cited papers are extremely different. They follow different citation curves, which may represent variations in the recognition function of the publications over time. Usually, the HCP curve is positively increasing with fewer fluctuations compared to RCP curves. The HCP starts dropping citations once a new related HCP raises. Furthermore, HCP has a wide variety of periodicals as well as papers from both related and unrelated subjects that frequently cite them. Figure 1 depicts an illustration of curve patterns for eight examples from HCP and RCP selected randomly from the collected dataset.

In Figure 1, we may observe the steady rise in citations received by HCP papers over time, while the RCP citations have fluctuated heavily during the same period. Another observation is the

DOI: 10.5530/jscires.12.1.008

number of citations received by HCP papers in a short period (One year from publication) is about ten times the number of citations received by RCP papers.

Numerous studies have looked into the elements of scientific articles that influenced how many citations were received. They discovered that an article having research collaboration received more citations overall than others.[6] Additionally, an article's initial citation and subsequent citations are closely associated.[2,7] Such findings lead us to think about detecting HCP papers from their bibliometrics information.

We continue our previous efforts to study the HCP papers using a small set of paper characteristics (16 features) as reported in GM BinMakhashen, *et al.*[8] the previous study was able to identify HCP articles with an accuracy of up to 89% using machine learning. In this study, we extended the extracted characteristics to 20 features and optimized the models to reach higher recognition rates as detailed in Section 5. Moreover, we utilized the Clarivate (Web of Science) database as an illustration. This study develops automatic HCP prediction using bibliometric characteristics. Later, the study findings can be applied to other research databases.

HCP Paper Criteria

Absolute and relative criteria are two ways to identify highly cited publications. These techniques have already been discussed in research[2] and utilized in previous research studies to select and examine publications for research excellence.[1,2,9,10] However, we built the dataset for this study using Clarivate Analytics' Essential Science Indicators (ESI). The selected paper is marked by the ESI as HCP if the paper received enough citations to be placed at the top 1% of its scientific field, otherwise, the selected paper is RCP. Due to its dynamic nature and connection to a specific academic topic, the ESI HCP thresholds are more frequently changing.

Research Contributions

In light of the observations made above, this study contribution is listed below:

1. Extending the feature set from 16 to 20 features, and,
2. Utilizing the initial citation information as indicated is important by other studies.[2,7]
3. Conducting thorough machine learning experiments to improve HCP recognition accuracy.
4. Comparing single ML models, Fusion of models, Ensemble models, and Deep Learning for HCP prediction.
5. Analytically highlights the importance of the HCP characteristics.

The paper is organized as follows: Section 2 presents previous related work; the background of the machine learning models is presented in Section 3; Then, the experimental methodology is discussed in Section 4; Section 5 presents and discusses the results; Finally, the study conclusions and future work is presented in Section 6.

Related Work

In this section, we discuss related research that studied highly cited papers' characteristics and machine learning for forecasting and prediction of HCP.

Highly Cited Papers

HCP can be characterized by its essential bibliometric information. These characteristics can be classified into three categories: paper, journal, and author. Among these characteristics are the title length, abstract length, and article pages.[11] Moreover, counts of keywords, references, tables, and Figures were also used for the HCP analysis.[12] Consequently, all these features are

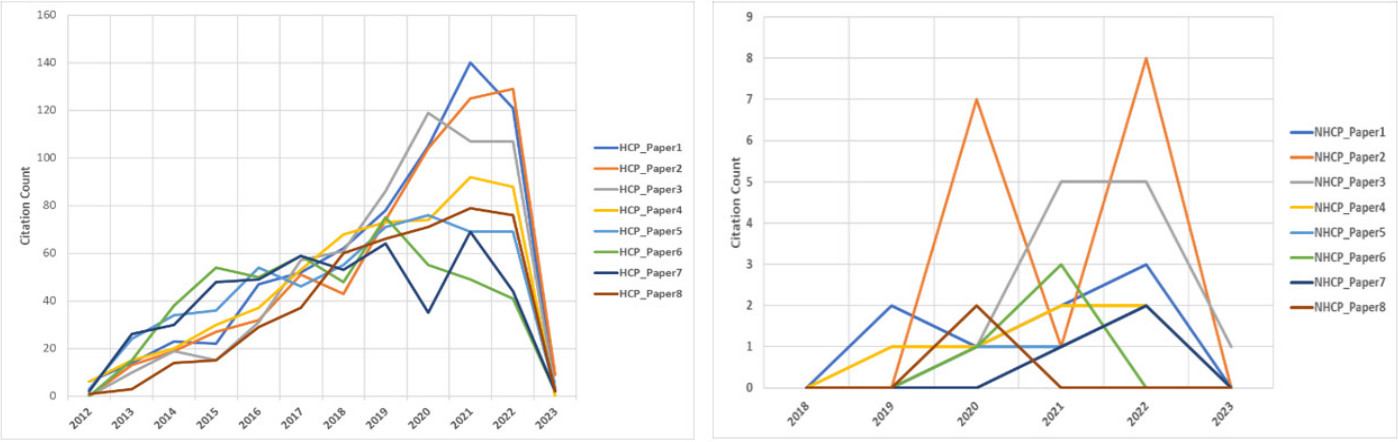


Figure 1: Eight citation patterns examples of published papers, A) High Cited Papers, B) Regular Cited Papers.

representing an essential characteristic that authors may consider to improve their articles' writing and get more citations.

Another important factor is the Journal Impact Factor (IF) or CiteScore (CS). A journal's impact can be viewed not only as a measure of journal quality but also as a measure of article quality. Therefore, the higher the IF or CS of a journal, the higher the confidence in the quality of its work. As a result, IF and/or CS contribute to the citation count of the paper.[12]

Several studies have shown strong statistical evidence that publication in high IF/CS journals is positively correlated with citation counts.[13,14] Moreover, the journal's coverage and scope are among the essential factors for HCP. For example, articles published in multidisciplinary journals are expected to receive more attention than other published papers in specialized journals.[14] Similarly, journals' coverage can play a positive or negative role in promoting the published articles to the journals' audience. Therefore, national journals may receive fewer citations than international journal counterparts.[15]

In general, the number of author characteristics had a great influence on the citation of the scientific papers. Noorhidawati *et al.*[16] noted that most of his HCP papers had between two and five authors, and 25 % of the papers had ten authors or more. Further studies found a positive correlation between the number of authors and the citation frequency.[11,17] In addition, other author-related factors such as authors academic rank, productivity, and reputation had explored.[17,18]

Different fields of study often have different citation thresholds, as calculated by WoS.[8] Noorhidawati *et al.*[16] reported that more than 50% of his HCP belonged to Technology and Engineering fields, with only 16% representing medicine. Another study found that works in the social sciences were cited more frequently than works in the natural sciences.[18] Additionally, citations can be affected by field size. For example, publications on organic chemistry, analytical chemistry, and physical chemistry are cited more frequently than those on biochemistry.[19,20]

Machine Learning Methods

There are two objectives to target HCP in the literature: forecasting future citation count, and early detection of next HCP paper. For citation forecasting, researchers use a set of features that represent paper, journal, and author information, current citation count etc. These features can be computed from samples of published research metadata.[21] Then, a regression-based methods are used to forecasting the future citation count.[10] By targeting citation count as a response variable, scientists may predict important phenomena such as breakthrough research, new areas of research, long-term scientific impact, etc.[22]

On the other hand, HCP detection allows researchers to early recognize recently published papers as the next HCP. Similar to the first objective, some feature should be extracted from the metadata and preprocessed to build machine learning models.[22]

Ponomarev *et al.*[10] reported a predictive model to predict future citation patterns using a time-dependent analysis of citation rates. Papers with high citations are indicating research excellence. However, this is not a sufficient condition for the paper to be considered groundbreaking research. Therefore, a multidimensional feature space was used.[23] Another work by Wang *et al.*[24] used 25 features to predict low, medium, and high citations. Their developed model provided projections for 15 years. In their work, first author, study quality, and journal reputation were found to be the most frequent predictors of citation frequency.

In addition, Wang *et al.*[22] integrated features extracted from both bibliometric and altimetric information to predict the rate of citation increase. Among the important factors identified was the leadership of the first author. Another study looked at early article citations for long-term citation regression.[25] However, long-term citations may not be maintained if groundbreaking research is identified. Thus, an early citation may not be effective in forecasting a long-term citation count. However, it could be used to build machine learning models with such initial information. The factors affecting the long-term citation growth of the paper were not fully identified.[26]

Contrary to the above results, Hurley *et al.*[27] reported that journal and linguistic characteristics were more important than the count of authors/co-authors in influencing citation behavior. They came to their conclusions using logistic regression models. In summary, identifying an effective set of characteristics can help researchers identify a list of tips to shape their scientific reports accordingly and draw attention to the findings their present.

In this study, we are developing a machine learning methodology for highly cited paper recognition using 20 features. Unlike the above studies where the task is to estimate the future citation count. In this study, we focus more on a set of HCP bibliometric information that can be utilized for HCP recognition. These findings will be of interest to writers and researchers in all disciplines seeking to improve their citation rates.

In this section, we present a brief background of each machine learning algorithm adopted in this study.

Support Vector Machines

SVMs are discriminant classifiers based on their assumptions that two types of data should be segregated.[28] So, the key task of SVM is to find a decision boundary that maximizes the segregation between the classes. This is done by optimizing the following equation:

$$\min_{w,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \text{s.t. } w^T$$

$$x_n t_n \geq 1 - \xi_n \forall_n \text{ and } \xi_n \geq 0 \forall_n \quad (1)$$

where $t \in \{-1; 1\}$ and ξ_n are penalties for those points that violate the decision margin. As a final step, the classification outcome is computed after selecting the best parameters and training the algorithm by:

$$\underset{t}{\operatorname{argmax}}(W^T X_{test})^t \quad (2)$$

For complex datasets, Cortes and Vapnik introduced a kernel trick in C Cortes, *et al.*[29] to transform the complex dataset into another easy-to-separate data by increasing dimensional space instantly. Therefore, Equation 2 is modified as follows:

$$\underset{t}{\operatorname{argmax}}(W^T \phi(X_{test}))^t \quad (3)$$

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

Logistic Regression

A logistic model (or logit model) is a statistical model that predicts the likelihood of an event occurring (i.e., A paper is HCP). Logistic regression is used in regression analysis to estimate the parameters of a logistic model. The event's log odds are a linear combination of one or more HCP characteristics. The logistic function is the Bernoulli distribution's natural parameter and the "simplest" way to convert a real number to a probability as in Equation 4.

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

where β_0 and β_1 are the model parameters to be optimized using likelihood maximizing estimation.

Naive Bayes

Probabilistic-based supervised learning algorithms such as the Naive Bayes algorithm are used for solving classification problems based on the Bayes theorem. A Naive Bayes Classifier is one of the simplest and most efficient algorithms for building machine learning models able to predict events quickly. In general, an object's probability is used in a probabilistic classifier to predict its likelihood. In Bayes' theorem, prior knowledge is used to determine the probability of a hypothesis. It is also known as Bayes' rule which has the following formula:

$$p(A|B) = \frac{\mathrm{P(B|A)}P(A)}{\mathrm{P(B)}} \quad (5)$$

Where $P(A)$ is the prior probability, $P(B)$is the marginal probability, $P(A|B)$ is the posterior probability, and P(B|A) is the likelihood probability.

K-Nearest Neighbor

kNN is nonparametric classifier that identifies a query label based on evidence from its neighborhood samples in the training set. Furthermore, since it is a nonparametric method, the algorithm makes no strong assumptions about the decision space before evaluation [30]. Formally, a query sample x defines a set of the k nearest neighbors of x as S_x, where,

$$S_x \subset D \; s.t. \left|S_x\right| = k \; and \; \forall (x', y') \in S_x.$$

$$dist(x, x') \leq \min_{\{(x'', y'') \in S_x\}} dist(x, x'') \quad (6)$$

Then, $kNN(.)$ is defined as a function that returns the frequent label of the samples in S_x:

$$kNN(x) = mode(y'' : (x'', y'') \in S_x \quad (7)$$

where mode(.) is returning the most frequent label in the S_x.

Decision Trees

Decision tree method classify data instances by estimating the most appropriate rule along the tree (i.e., from the root node to the leaf node) to enable such algorithms to make decisions. The algorithm refines the decision for specific characteristics of the instance by iteratively choosing a branch/sub-decision at each point (node). When the algorithm reaches a leaf node, it produces a final decision (class label). Intermediate nodes can have at least two branches (child nodes), but leaf nodes have no children.

The decision tree algorithm is simple, but due to its structure, the number of nodes can be very large. In this work, decision trees are constructed using the Gini impurity method. Assuming we have C classes and p(i) is the probability of choosing an instance with class label *i*, the Gini impurity is computed as:

$$G = \Sigma^C p(i) \times (1 - p(i)) \quad (8)$$

This is computed for all possible classes, and the smaller the Gini value is the one selected for splitting the branches of the decision tree.

Random Forest

Random Forest is a well-known ensemble algorithm for machine learning. This algorithm uses the bootstrap aggregation technique (bagging)[31] to build multiple uncorrelated decision trees. So, as the name suggests, a random forest builds a large number of individual decision trees that act as an ensemble classifier. Each tree in the random forest refines the rules for class prediction. Then the most frequently selected label from all trees becomes the class label.

Multilayer perceptron Neural Networks

A fully connected class of feedforward Artificial Neural Networks (ANN) is called a Multilayer Perceptron (MLP). It has been confirmed that the term "MLP" has been used ambiguously. It is sometimes applied broadly to any feedforward ANN, and at other times it is applied specifically to networks made up of several layers of perceptrons (with activation functions). The MLP architecture consists of three main stages; input, hidden,

and output layers. In the input layer, the data features are simply fed to the network with the forward-connected network to the hidden layer. At each hidden layer, several neurons are firing with certain activation functions. Then, the last group of neurons in the hidden layer is connected to the output layer. The output layer has a certain number of neurons equal to the number of expected outputs. Usually, the output layer produces a score value that can be interpreted as a probability to determine the label of the input sample as a positive class if it is greater than a certain threshold, otherwise, it is a negative class sample. Figure 2 depicts the MLP architecture. Such ANN network requires training to tune the internal and output models' parameters. Such training is carried out using a backpropagation algorithm as described in R Rojas, *et al*.[32] The main function of the backpropagation algorithm is to converge to the optimal model's parameters. The ANN network utilizes the learning error to control the speed of model convergence.

Long short-term memory networks

An artificial neural network called Long Short-Term Memory (LSTM) is employed in deep learning and artificial intelligence. LSTM has feedback connections as opposed to typical feedforward neural networks. Such a Recurrent Neural Network (RNN) can process entire data sequences in addition to single data points such as images, speeches, or videos. Because of this feature, LSTM networks are perfect for handling and forecasting data.

A standard RNN is analogized to have both "long-term memory" and "short-term memory" in the name of the LSTM. The activation patterns in the network change once per time step, analogous to how physiological changes in synaptic strengths store short-term memories. The connection weights and biases in the network change once per episode of training, analogous to how physiological changes in synaptic strengths store long-term memories. Figure 3 illustrates the LSTM single-cell structure. As we can observe the neuron shows an addition to a computation unit, the neuron has a state unit and short and long memories decision gates. The input to this neuron is also combined with the previous output of the cell (recurrent) to improve the neuron's future output.

METHODOLOGY

In this section, we discuss the setup of the experiments by describing the dataset and machine learning models' configuration.

Data collection

The database used in this paper covers both highly and regularly cited papers that were retrieved by performing the Clarivate query (WS=(Technology)) that was published between 2009 to 2022. This query includes publications from eight research fields as categorized by Clarivate - Web of Science. These research domains are listed below:

1. Construction and Building Technology
2. Food Science and Technology
3. Green, Sustainable Science and Technology
4. Imaging Science and Photographic Technology
5. Medical Laboratory Technology
6. Nuclear Science and Technology
7. Quantum Science and Technology
8. Transportation Science and Technology.

The retrieved data was filtered to focus only on Article or Review types. The total dataset is 32700 records and there are 10231 HCP records with more than 1,503,484 citations. We extracted bibliometric information from each record as tabulated in Table 1. We grouped them into three categories of features; 1) Author-based features represents the number of authors, interdisciplinary and international/national collaboration,

Table 1: Features Definitions.

Category	Feature	Definition
Author-based	AU	Count of authors
	COL	Inter. Collaboration : Count of Countries
	AFF	Interdisciplinary: Count of Affiliations
Article-based	FUN	Research Funding (True, False)
	DT	Title Length
	TP	Title punctuation count
	KW	Count of Keywords (author)
	KWP	Count of Keywords (Clarivate)
	ABS	Abstract length of characters
	REF	References count
	YEAR	Publication year
	CCW	Citation Count WoS
	CCT	Citation Count all databases
	USAGE180	Paper usage in the last 180 days
	USAGET	Total paper usage to date
Journal -based	VOL	Journal Volume
	ISSUE	Journal Issue
	SP	Special Issue (True, False)
	RACOUNT	Count of Research Areas
	OA	Open Access (True , False)

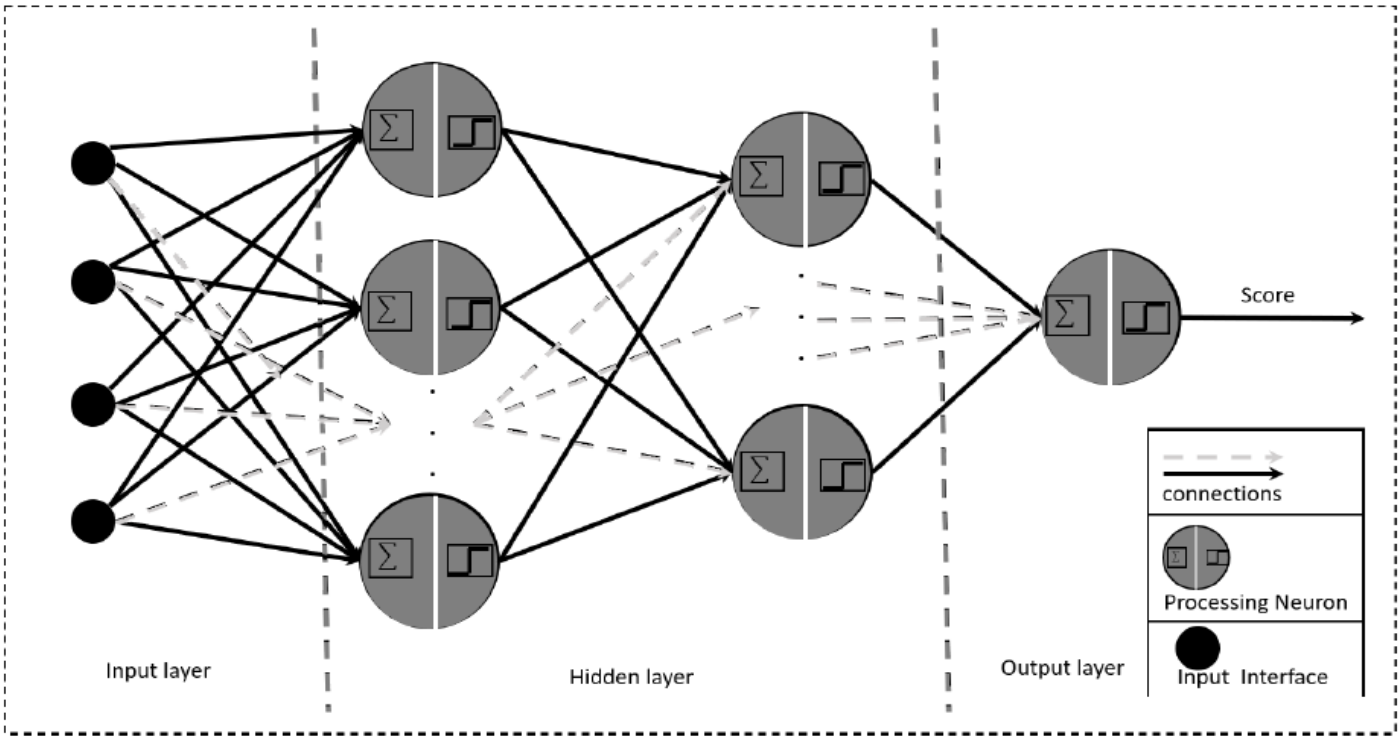


Figure 2: Multilayer Perceptron Neural Network Architecture.

2) Article-based characteristics is including bibliometric information and text-based analysis such as the length of the title, number of punctuations, number of words in the abstract, etc., and 3) journal-based features cover the age of a journal (volume), issues, how interdisciplinary the journal is?, and whether the paper is published in an open-access style or not etc.

AI Models Configuration

Eight ML models were developed and analyzed in this study. The default parameters were found suitable for our dataset or the model is limited to fixed configurations such as Logistic Regression and Naive Bayes models. The Multilayer Perceptron classifier is set to its default setting.[33] Moreover, Decision trees and Random Forests were configured with Gini impurity method.

KNN is among the simple ML methods, yet, it can produce very complex decision space. Determining the K value is very important and usually, determined heuristically by an elbow method. Figure 4 illustrates KNN performances using various K values. The highest performance was recorded using K = 1, however, the decision space of such a k value is complex. Therefore, K was set to 3 for the rest of the experiments.

Support Vector Machines algorithm has several hyper-parameters to tune. We can set how resilient SVM should be with violating samples (misclassification). This hyper-parameter allows us to develop flexible decision boundaries. Moreover, the data may not be separable, we need to configure the SVM with the right kernel to overcome such a problem. each of these kernels (except the liner) has several other parameters to tune. To automate such hyper-parameter selection, the Bayesian optimization method was adopted to find the best SVM hyper-parameters. Therefore, SVM hyper-parameters were set to C =10.0, degree=2, gamma=10.0, with a polynomial kernel.

LSTM is constructed with a shallow structure using Keras sequential model. The model has two layers of LSTM at the input stage. With 128 and 32 neurons. These two layers have 96,896 trainable parameters. To regularize the LSTM network Dropout layers are introduced to avoid overfitting issues. These two layers are injected between the LSTM layers to the first Dense layer. The Dense layers are ANN network layers similar to MLP. The last Dense layer consists of one neuron to determine the label of the input. These two Dense layers have a total of 2177 trainable parameters. Table 2 lists the structure of the LSTM model.

RESULTS AND DISCUSSION

Feature Analysis

In this section, we discuss the data feature space. Table 1 tabulates the descriptive information of the dataset. The information represents the raw data as we can notice from the stats. The mean of authors per paper is about five authors per paper in general. However, HCP has a mean of six authors per paper and four authors for RCP. By studying the initial citations and usage, the mean citation of 187 citations for HCP and about 26 citations for RCP. RCP is accessed at least 7 times on average, while the HCP was used at least 30 times in the last 180 days. The complete descriptive stats of the data are tabulated in Table 3.

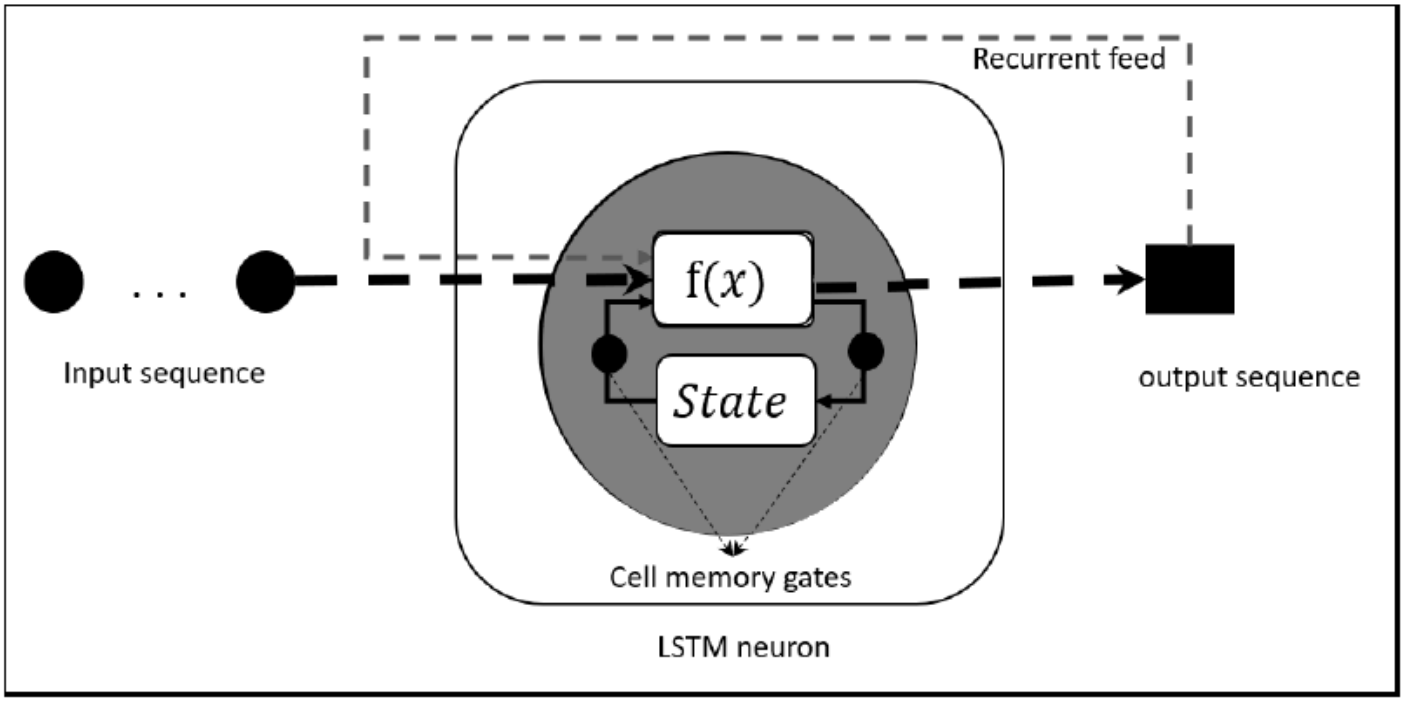


Figure 3: Long Short Memory Networks: The complexity of an LSTM Neuron.

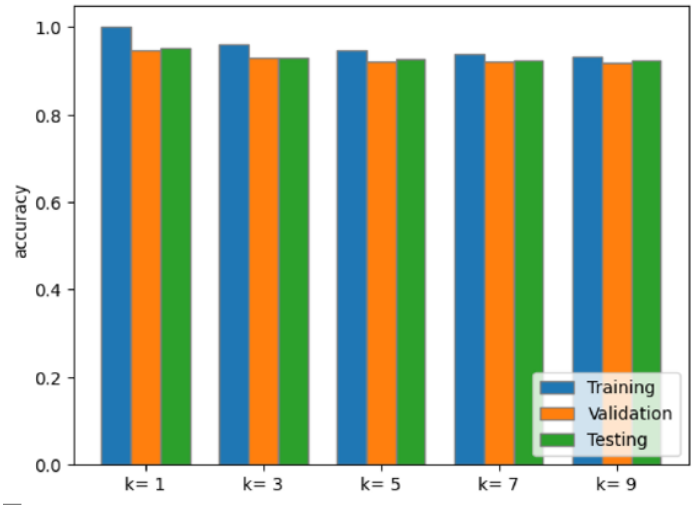


Figure 4: KNN Model Selection.

The skewness and Kurtosis of the data distribution are also computed. The skewness represents the shape of the distribution. It can be quantified to define the extent to which a distribution differs from a normal distribution. Almost all features are skewed Figure 5A. Most of these features are right-skewed while DT, KW, KWP, RACOUNT, SP, and OA are left-skewed.

Moreover, we calculated the kurtosis of the dataset to understand how thick these skewed tails are. In other words, Kurosis measures the thickness or/and heaviness of the distribution. Kurtosis may compute the height of the data distribution. Figure 5B confirms thick tails for most of the characteristics as indicated with extremely positive values.

Table 2: LSTM Model Structure (99,073 Total Trainable Parameters).

Layer (type)	Output Shape	Param #
lstm_71 (LSTM)	(None, None, 128)	76,288
lstm_72 (LSTM)	(None, 32)	20,608
dropout_117 (Dropout)	(None, 32)	0
dense_110 (Dense)	(None, 64)	2,112
dropout_118 (Dropout)	(None, 64)	0
dense_111 (Dense)	(None, 1)	65

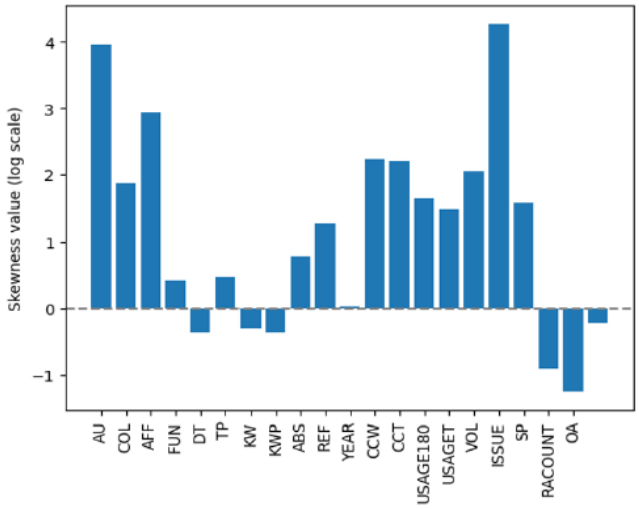
Experimental Results and Discussion

Table 4 summarizes the experiment results. The results illustrated in Table 4 is showing that the ensemble methods outperformed other ML and DL models in our experiments. Using the Random Forest algorithm, the model was able to generalize and predict HCP papers correctly with 97.9% accuracy using 20 features. The ensemble method created complex decision boundaries flexible enough to recognize HCP papers. Compared with the Logistic Regression (LR) classifier, the latter is limited with a separable dataset and it is a known limitation of LR to find a clear decision boundary in the non-separable datasets.

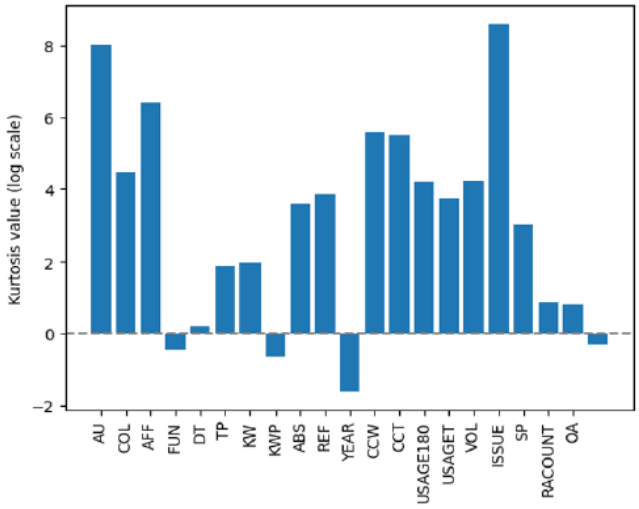
A second good algorithm for the study dataset is a decision tree algorithm based on the performance reported in Table 4. Given the complexity of the algorithm, the Decision Tree algorithm is less complex than Random Forest which gives the Decision Tree the preference. Moreover, the training time for Decision Trees was 26 times faster than Random Forests Table 5.

Table 3: Dataset descriptive statistics.

Feature	Mean	Std	Min	25%	50%	75%	Max
AU	5.19	17.31	1	3	4	6	1192
COL	1.71	1.50	0	1	1	2	42
AFF	3.31	5.12	0	1	2	4	248
FUN	0.66	0.48	0	0	1	1	1
DT	12.64	4.43	1	10	12	15	51
TP	1.40	1.30	0	1	1	2	23
KW	5.10	2.55	0	4	5	6	38
KWP	8.04	3.30	0	7	10	10	10
ABS	197.45	67.41	0	153	194	234	2088
REF	103.08	72.47	0	54	88	134	2030
YEAR	2018	3.15	2009	2017	2019	2021	2022
CCW	74.53	135.81	0	6	25	92	6864
CCT	77.08	140.67	0	6	25	95	6919
USAGE180	14.73	22.02	0	2	7	19	684
USAGET	94.91	146.50	0	15	40	113	3458
VOL	81.77	210.36	0	11	29	80	2676
ISSUE	5.16	29.58	0	0	2	8	2493
SP	0.04	0.19	0	0	0	0	1
RACOUNT	2.11	0.84	1	2	2	3	5
OA	1.24	0.87	0	1	1	2	4



(a) Skewness of the features



(b) Kurtosis of the features

Figure 5: Features Skewness and Kurtosis.

Other algorithms (SVM, MLP, KNN, and Naive Bayes) achieved similar performances in Train/Testing evaluation setting. In this setting, the data was divided into training and testing sets randomly. Such evaluation settings may maybe deceive researchers on model performance. As all these models achieved high performance in this setting, the Naive Bayes algorithm suffered significantly in the K5Fold evaluation setting. We can observe that the Naive Bayes achieved 86.2% average accuracy.

We conducted an experiment to assess the models' integration (i.e., fusion). The integration of models was carefully conducted. We set weights relative to their single-shot performance and set the highest weight to the Random Forest classifier. Table 5 lists the weights of the training time of each model.

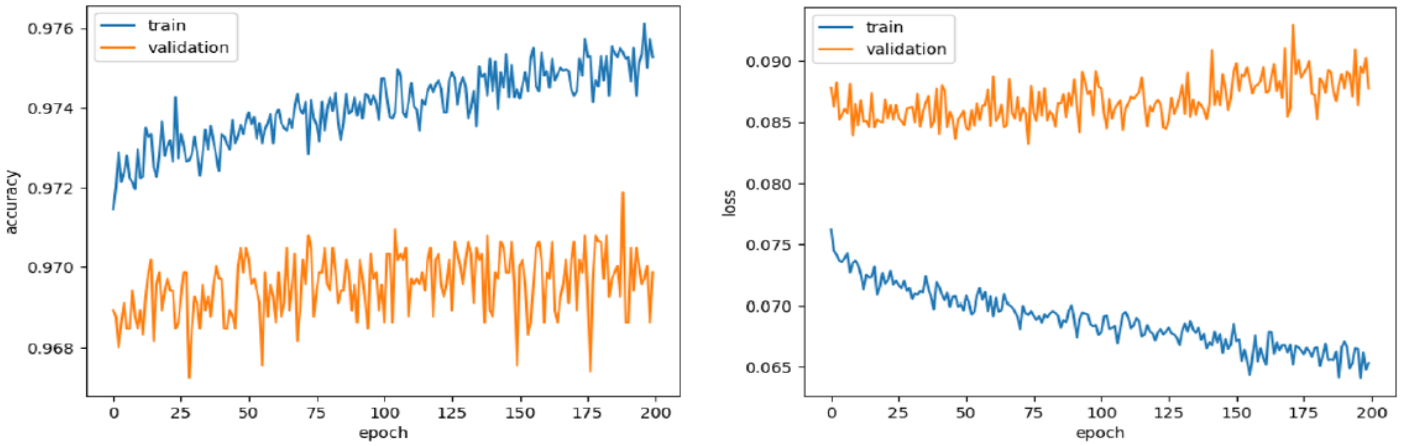


Figure 6: LSTM Model Training History.

Table 4: Experiments Results: Performance Accuracy.

Model	Train/Test	K5Fold
Logistic	87.61%	87.46%
Decision Tree	96.71%	96.07%
KNN	94.70%	93.75%
SVM	91.99%	91.82%
Random Forest	97.91%	97.50%
MLP	95.55%	95.54%
Naïve Bayes	95.54%	86.24%
Model Fusion	97.44%	97.57%
LSTM	96.31%	95.39%

Table 5: Model Fusion Weights and Training Time.

Model	Weights	Training Time
Logistic Regression	0.05	0.6 Sec
Decision Tree	0.121	0.7 sec
KNN	0.125	0.0 sec
Naïve Bayes	0.12	0.0 sec
SVM	0.126	4.3 min
MLP	0.126	2.5 min
Random Forest	0.33	15.6 Sec

Finally, with 200 epochs to build the LSTM model the training and validation started with 96% to 97% accuracy and fluctuated around the convergence minimum loss. Figure 6 illustrates model training accuracy and losses. We notice that the validation loss is slightly increased which may indicate the start of overfitting. The results of LSTM achieved high performance with 96.3% accuracy in Train/Test setting. Tabular data is claimed to be not suitable for deep learning models,[34] yet such a topic is still under investigation.

CONCLUSION

Highly cited papers are important indicators of research excellence in various fields. Usually, researchers and research management track such HCP to keep on top of their research domains and generate state-of-the-art research ideas. In this study, we extracted 20 characteristics from Clarivate papers' records to build an automatic machine-learning model capable of recognizing HCP with high accuracy. We notice that the set of characteristics extracted in this paper is significant in representing HCP and RCP papers. The tree-based ensemble models were the best to recognize HCP in tabular data format. Moreover, other models with complex decision boundaries can also achieve such high performance. A simple Decision tree algorithm would be the best as it was built in 0.6 sec using 26,160 training records.

Future research will compare the proposed methodology to Clarivate using multiple databases, including Elsevier-SCOPUS and Dimensions. For research management and users, estimating the overlap among these paid databases would be crucial. Additionally, features and characteristics can be automatically extracted for deeper and more complex models using natural language processing (NLP) techniques.

ACKNOWLEDGEMENT

The authors would like to acknowledge the help and support provided by King Fahd University of Petroleum and Minerals (KFUPM).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ABBREVIATIONS

ANN: Artificial Neural Networks; **HCP:** Highly Cited Papers; **RCP:** Regularly Cited Papers; **KNN:** K-Nearest Neighbors; **SVM:** Support Vector Machines **WoS:** Web of Science; **LSTM:** Long-Short Term Memory; **MLP:** Multilayer Perceptron.

REFERENCES

1. Persson O. Are highly cited papers more international? Scientometrics. 2010;83(2):397-401.
2. Aksnes DW. Characteristics of highly cited papers. Research Evaluation. 2003;12(3):159-70.
3. Van Raan AF. Citation analysis: Measuring scientific. The Web of knowledge: A Festschrift in Honor of Eugene Garfield. 2000:301.
4. Zeng A, Shen Z, Zhou J, Wu J, Fan Y, Wang Y, *et al*. The science of science: From the perspective of complex systems. Physics Reports. 2017;714-715:1-73.
5. Catalini C, Lacetera N, Oettl A. The incidence and role of negative citations in science. Proceedings of the National Academy of Sciences. 2015;112(45):13823-6.
6. Velez-Estevez A, García-Sánchez P, Moral-Munoz JA, Cobo MJ. Why do papers from international collaborations get more citations? a bibliometric analysis of library and information science papers. Scientometrics. 2022;127(12):7517-55.
7. Winnink JJ, Tijssen RJW. Early-stage identification of breakthroughs at the interface of science and technology: Lessons drawn from a landmark publication. Scientometrics. 2015;102(1):113-34.
8. BinMakhashen GM, Al-Jamimi HA. Evaluation of machine learning to early detection of highly cited papers. In 2022 7th International Conference on Data Science and Machine Learning Applications. IEEE. 2022:1-6.
9. Antonakis J, Bastardoz N, Liu Y, Schriesheim CA. What makes articles highly cited? The Leadership Quarterly, 2014;25(1):152-79.
10. Ponomarev IV, Williams DE, Hackett CJ, Schnell JD, Haak LL. Predicting highly cited papers: A method for early detection of candidate breakthroughs. Technological Forecasting and Social Change. 2014;81:49-55.
11. Bornmann L, Leydesdorff L, Wang J. How to improve the prediction based on citation impact percentiles for years shortly after the publication date? Journal of Informetrics. 2014;8(1):175-80.
12. Elgendi M. Characteristics of a highly cited article: A machine learning perspective. IEEE Access. 2019;87(7):977-86.
13. Didegah F, Thelwall M. Which factors help authors produce the highest impact research? Collaboration, journal and document properties. Journal of Informetrics. 2013;7(4):861-73.
14. Vanclay JK. Factors affecting citation rates in environmental science. Journal of Informetrics. 2013;7(2):265-71.
15. Millet-Reyes B. The impact of citations in international finance. Global Finance Journal. 2013;24(2):129-39.
16. Noorhidawati AYIA, Zahila MN, Abrizah A. Characteristics of Malaysian highly cited papers. Malaysian Journal of Library and Information Science. 2017;22(2):85-99.
17. Biscaro C, Giupponi C. Co-authorship and bibliographic coupling network effects on citations. PloS One. 2014;9(6):e99502.
18. Chakraborty T, Kumar S, Goyal P, Ganguly N, Mukherjee A. Towards a stratified learning approach to predict future citation counts. In IEEE/ACM Joint Conference on Digital Libraries. IEEE. 2014;351-60.
19. Bornmann L, Schier H, Marx W, Daniel HD. What factors determine citation counts of publications in chemistry besides their quality? Journal of Informetrics. 2012;6(1):11-8.
20. Gallivan MJ. Analyzing citation impact of is research by women and men: Do women have higher levels of research impact? In Proceedings of the 50th Annual Conference on Computers and People Research. 2012:175-84.
21. Tahamtan I, Safipour AA, Ahamdzadeh K. Factors affecting number of citations: A comprehensive review of the literature. Scientometrics. 2016;107(3):1195-225.
22. Wang F, Fan Y, Zeng A, Di Z. Can we predict esi highly cited publications? Scientometrics. 2019;118(1):109-25.
23. Ponomarev IV, Lawton BK, Williams DE, Schnell JD. Breakthrough paper indicator 2.0: Can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction? Scientometrics 2014;100(3):755-65.
24. Wang M, Yu G, Xu J, He H, Yu D, An S. Development a case-based classifier for predicting highly cited papers. Journal of Informetrics. 2012;6(4):586-99.
25. Abrishami A, Aliakbary S. Predicting citation counts based on deep neural network learning techniques. Journal of Informetrics. 2019;13(2):485-99.
26. Yan R, Tang J, Liu X, Shan D, Li X. Citation count prediction: learning to estimate future citations for literature. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management. 2011:1247-52.
27. Hurley LA, Ogier AL, Torvik VI. Deconstructing the collaborative impact: Article and author characteristics that influence citation count. Proceedings of the American Society for Information Science and Technology. 2013;50(1):1-10.
28. Hsu CW, Lin CJ. A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks. 2002;13(2):415-25.
29. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20(3):273-97.
30. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 1992;46(3):175-85.
31. Liaw A, Wiener M, *et al*. Classification and regression by randomforest. R News. 2002;2(3):18-22.
32. Rojas R. The backpropagation algorithm, in Neural networks. Springer. 1996:149-82.
33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: Machine learning in Python," Journal of Machine Learning Research. 2011;12:2825-30.
34. Fayaz SA, Zaman M, Kaul S, Butt MA. Is deep learning on tabular data enough? an assessment. International Journal of Advanced Computer Science and Applications. 2022;13(4).

Cite this article: BinMakhashen GM, Al-Jamimi HA. Intelligent Prediction of Next Highly Cited Paper using Machine Learning. J Scientometric Res. 2023;12(1):44-53.