Scientometrics (2022) 127:849–869
https://doi.org/10.1007/s11192-021-04240-2

Research trend prediction in computer science publications: a deep neural network approach

Soroush Taheri[1] · **Sadegh Aliakbary**[1]

Received: 7 February 2021 / Accepted: 26 November 2021 / Published online: 20 January 2022

Abstract
Thousands of research papers are being published every day, and among all these research works, one of the fastest-growing fields is computer science (CS). Thus, learning which research areas are trending in this particular field of study is advantageous to a significant number of scholars, research institutions, and funding organizations. Many scientometric studies have been done focusing on analyzing the current CS trends and predicting future ones from different perspectives as a consequence. Despite the large datasets from this vast number of CS publications and the power of deep learning methods in such big data problems, deep neural networks have not yet been used to their full potential in this area. Therefore, the objective of this paper is to predict the upcoming years' CS trends using long short-term memory neural networks. Accordingly, CS papers from 1940 and their corresponding fields of study from the microsoft academic graph dataset have been exploited for solving this research trend prediction problem. The prediction accuracy of the proposed method is then evaluated using RMSE and coefficient of determination (R^2) metrics. The evaluations show that the proposed method outperforms the baseline approaches in terms of the prediction accuracy in all considered time periods. Subsequently, adopting the proposed method's predictions, we investigate future trending areas in computer science research from various viewpoints.

Keywords Scientometrics · Research trends · Time-series prediction · Deep learning · Computer science

Introduction

With the advent of theoretical science in the past century, the number of published research papers has grown significantly. Moreover, with the prevalence of the internet, these scholarly works are made available to a broad range of researchers. Investigating these research works and making predictions on their directions and trends is not a trivial task and cannot

✉ Sadegh Aliakbary
s_aliakbary@sbu.ac.ir

Soroush Taheri
so.taheri@mail.sbu.ac.ir

[1] Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran

be scientifically performed without adopting data-driven methods (Clauset et al., 2017). Therefore, an exclusive research area known as Scientometrics, or the science of science (Fortunato et al., 2018), has emerged to analyze all these scientific papers quantitatively using the available data (Leydesdorff, 2001).

Estimations showed that at least 247 million papers were available on the web at the end of 2020;[1] tens of thousands more being added to them daily. This vast number of researches being done has created a need for big data approaches in Scientometric works, also known as Big Scholarly Data (BSD) (Xia et al., 2017; Dridi et al., 2020). Since these studies can be beneficial to the whole research community, large digital libraries, academic social networks, and academic search engines have been developed for easier access to this valuable data. Applying this data, researchers have published a considerable number of articles on introducing novel methods to investigate the most influential papers and solve academic recommendation, expert finding, and research trend analysis and prediction problems in different fields of study (Mahalakshmi et al., 2017; Xia et al., 2017).

Due to its significant growth in the past decades, Computer Science (CS) has always been one of the hottest research fields worldwide. Advances in computational technologies and the fact that other fields are becoming computational and are more data-driven than ever have made this happen. Therefore, many scholars, research institutions, universities, and research funding organizations are concentrating on CS scholarly works (Hoonlor et al., 2013). Considering all these people, institutions, and the extensive amount of funding on the CS research, knowing which branches of these studies are trending can be favorable for all parties. The scholars can use this knowledge to set a better career path, and the institutions, organizations, and universities can also define better policies for accepting students and researchers and granting more funds to trending research fields (Goodall, 2006). As a result, several scientometric papers have aimed to find the most trending research areas of CS. There are two different approaches regarding trends of research: the first approach is analyzing the current trending fields of study. In contrast, the second approach is trying to predict future trends, focusing on which research areas will become more prevalent in the upcoming years. Some of these studies concentrate on all of the CS research areas, while the others focus on subsets of a particular field. It is noteworthy that trends might have different definitions in this context. A trending field could be one with a vast number of research papers published in, or a research area that receives the most citations among all the other areas, or a field of study which has the most significant growth in terms of the number of publications or citations (Effendy and Yap, 2017).

In this paper, we concentrate on predicting research trends in Computer Science publications in the five upcoming years, using the number of publications in each field of study per year as the measure of trendiness. A Long Short-Term Memory (LSTM) deep neural network is proposed for predicting the number of publications in each field and is able to predict the trends in the primary CS research areas as well as their subsets.

The remaining of this paper is organized as followed: in "Related works" section, the state-of-the-art works are reviewed. "Proposed method" section includes the problem statement and a precise definition of the proposed method. In "Evaluations" section, the applied dataset and the baseline methods are introduced, and the proposed method's performance is evaluated against them. "Discussion and results" section contains the results of this

[1] https://academic.microsoft.com/publications.

study. Finally, in "Conclusion" section, we conclude this paper's achievements and discuss the possible future works.

Related works

There are different perspectives on trends in state-of-the-art papers of the research trend studies. The trendiness can be defined in terms of the number of published papers in each field of study (general trend), or it can be defined by the number of citations each research area receives altogether (citation trend) (Effendy and Yap, 2017). In contrast, while some research areas do not meet these two criteria, their significant growth shows their trendiness (emerging trends). It is essential to note that these three types of trends are distinct. Effendy et al. showed that in most cases, the citation trends present more fundamental fields of study, which gain more citations due to their nature (Effendy and Yap, 2017). Some other fields of study do not receive enough publications or citations to be considered general trends or citation trends; instead, they demonstrate an extraordinary growth in the number of their research works (Salatino et al., 2018).

Some of the important research works on the first perspective, namely general trends, are introduced in the following parts of this section, while the other types of studies are not explained in detail since they are irrelevant to our proposed method. Some of the investigated studies in this section only analyze and announce the current trending fields, while others further establish a method to predict future trends. Several non-CS papers are also introduced due to the importance and novelty of their methods.

Hoonlor et al. (2013) and Ebadi et al. (2020) analyzed trends concentrating on the relationship between trends and research funds. Wu et al. (2016) focused on the top 1% of active scholars as representatives of the population and studied their collaborations on co-authorship networks using Clauset–Newman–Moore algorithm (Clauset et al., 2004). Garousi and Ruhe (2013) studied regional trends on Software Engineering (SE) publications, comparing the effects of different countries on the SE trending subdomains. In another study, Wang et al. (2014) examined the top 20 SE trends based on macro-keywords (collected from classifications) and micro-keywords (based on user-defined topics) by utilizing the score generated to weigh each scholarly work's quality. Cheng et al. (2015) offered a rule-based anomaly detection model, namely WSARE, to investigate the shifting of trends from traditional domains to modern ones using network evolution analysis on topic co-appearance graphs. Likewise, Katsurai and Ono (2019) proposed a research trend mapping approach to detect rapid developments in dynamic co-word networks' edge weights. Some other studies analyzed the relationship between different types of trends. For instance, Pham et al. (2011) and Effendy and Yap (2017) took advantage of citation network structures and analyzed trends in interdisciplinary areas and the relationships between citation trends and general trending domains.

Yuen-Hsien Tseng et al. (2009) compared six different metrics on varying periods and duration for predicting trends in safety agriculture and information retrieval fields. They suggested API (Average Percentage of Increase) for large year spans and SLP (Slope of Linear Regression) for the rest of the cases. Sari and Widodo (2012) proposed an Extreme Learning Machine (ELM) to improve the drawbacks of feed-forward neural networks, then predicted the number of publications on IEEE research papers. Hurtado et al. (2016) used association rule mining to extract topics from four prestigious data mining and machine learning conferences. They developed a linear regression ensemble forecasting model to

take the effects of random other fields of study into consideration while predicting the number of publications of a target *Field of Study (FoS)*. The experiments demonstrate that considering six random fields results in the most accurate predictions. The ultimate prediction is the mean of 100 iterations of forecasting results to predict one forthcoming year's trends in this study. This paper is used as one of our baseline methods for comparison. Jabłońska-Sabuka et al. (2014) forecasted filtration and rectification trends of Chemical Engineering researches, proposing a MATLAB animal population dynamics model after co-occurred keywords were clustered using Sitarz and Kraslawski's method (Sitarz and Kraslawski, 2017). Instead of using the number of publications, Behrouzi et al. (2020) used five machine learning link prediction algorithms (Random Forest Classifier, Support Vector Machine, K-Nearest Neighbors, Multinomial Naive Bayes, and Gaussian Naive Bayes) on keyword networks to anticipate their future structures. Chen et al. (2018) proposed the Correlated Neural Influence Model (CONI) - a Gated Recurrent Unit RNN - to integrate the influences of top Artificial Intelligence and Data Mining conferences on one another and compared the results against three baseline methods (DeGroot, 1974; Hegselmann et al., 2002; Brockwell and Davis, 2013) and one naive predictor named 'Last Year' to demonstrate that CONI outperforms them all. In another research, Krenn and Zeilinger (2020) built a semantic network of concepts in which the links demonstrate that concepts are correlated and mentioned in the same research articles. An artificial neural network is applied for predicting the trending concepts and their divergence in the next 5 years in quantum physics' studies. Likewise, Rzhetsky et al. (2015) investigated network characteristics and the growth of a knowledge network, in which degree centrality corresponds to a node's importance. They provided a quantitative approach to understand research strategies and directions from the network data by formulating researchers' behavior on topic selection. Table 1 summarizes the trend prediction approaches analyzed in this section.

It is observed that despite the advances in artificial-intelligence-based methods to interpret time-series data and perform regression predictions, they are not yet utilized to their full potential in research trend prediction problems. Most machine learning or deep learning solutions concentrate on the articles' text to recognize the trending areas or use more traditional learning algorithms such as support vector machines to perform the regression predictions. More specifically, the capability of modern deep learning approaches to analyze sequential numerical values independently is not yet utilized to its full potential in this research domain. Therefore, in this study, we intend to take advantage of the Long Short-Term Memory algorithm as one of the best-performing deep learning approaches for sequential data analysis to forecast research trending areas.

Proposed method

In this section, we define the specific problem at hand as well as our adopted presumptions. Thereafter, we introduce the proposed method and its implementation details for solving the stated problem.

Problem statement

Assuming that the the actual number of publications for the target field of study f in the year k is referred to as p_f^k, its predicted value is defined as $\hat{p}_f^k$. Using the number of published papers in the past 10 years for any target field of study, namely p_f^y, p_f^{y+1}, ... , p_f^{y+9}, we

Table 1 Summary of the key trend prediction methods proposed in recent years

Year	Paper	Trend criteria	Method	Domain
2009	Yuen-Hsien Tseng et al. (2009)	API (Average Percentage of Increase) on number of publications	Statistical Formulation	Safety Agriculture/Information Retrieval
2012	Sari and Widodo (2012)	Number of publications (10 y to 1 y)	Extreme Learning Machine (ELM)	Computer Science
2014	Jabłońska-Sabuka et al. (2014)	Keywords co-occurrence clustering with Sitarz et al.'s method (Sitarz and Kraslawski, 2017)	Animal Population Dynamics Model	Chemical Engineering
2015	Rzhetsky et al. (2015)	Degree centrality in a knowledge network	Statistical Formulation	Biomedical Chemistry
2016	Hurtado et al. (2016)	Number of publications (9 y to 1 y) + 6 random other fields of study	Linear Regression Ensemble Forecasting	Data Mining/Machine Learning
2018	Chen et al. (2018)	Number of publications on top conferences	Gated Recurrent Unit (GRU) RNN	Artificial Intelligence / Data Mining
2020	Krenn and Zeilinger (2020)	Link weights on a semantic network	Artificial Neural Network (ANN)	Quantum Physics
2020	Behrouzi et al. (2020)	Links on keyword networks	Comparison of Multiple Machine Learning Methods for Link Prediction	Computer Science

Most of these research works consider the number of published papers or the directions and growth of topic networks as the trendiness criteria

aimed to predict the five upcoming years' count of publications p_f^{y+10}, p_f^{y+11}, p_f^{y+12}, p_f^{y+13}, and p_f^{y+14}. Therefore, the problem is to predict the five future years' publication count for any given field of study (FoS), having a time-series of the past 10 years of its number of publications on a yearly basis. This time-series is the only data taken into account, and no other features are utilized to solve this problem. Equation 1 formulates the problem for any given FoS.

$$\forall f \in \mathbb{FOS} : \text{given } [p_f^{y}, p_f^{y+9}] \xrightarrow{\text{predict}} [\hat{p}_f^{y+10}, \hat{p}_f^{y+14}] \tag{1}$$

The objective of this research is to perform such forecasting as well as minimizing the corresponding prediction error. Furthermore, a satisfactory method must be able to accurately predict both the yearly and the cumulative number of publications in the forthcoming years. Therefore, we also consider the total number of published papers in the five upcoming years. The actual cumulative number is specified as P_f, while the aggregated predicted value for these years is called $\hat{P}_f$. Consequently, the prediction accuracy of the proposed method will be evaluated by all $\hat{P}_f$ and $\hat{p}_f^k$ values.

Publication trend prediction

As stated in the introduction section, we have taken advantage of a machine learning approach to predict the number of publications, therefore predicting the general trends of variant CS topics using a time-series of the past 10 years of publication data. This time-series prediction falls into the regression prediction problems category, in which the aim is to provide a non-negative number in the output according to the input. It is proved that the machine learning approaches show better performance than the human-designed programs in these data-driven problems (Han et al., 2011). Among such machine learning methods, Artificial Neural Network (ANN) models are suitable solutions to overcome the complexity of regression predictions' non-linear behavior (Goodfellow et al., 2016). ANNs are used in a vast range of studies including computer vision (Wang and Sng, 2015), speech (Deng et al., 2013), face and handwriting recognition (Poznanski and Wolf, 2016), and natural language processing (Young et al., 2018), as well as more practical domains such as smart cities (Wang and Sng, 015). They are also applied to regression-only problems such as stock market analysis (Chong et al., 2017; Dargan et al., 2019). It is also shown that in more complex problems where a large amount of data is available, deeper neural networks with more hidden layers are able to outperform the more shallow ones, resulting in a new trending area known as *"Deep Neural Networks (DNN)"* or *"Deep Learning"* (Goodfellow et al., 2016).

Accordingly, we have proposed a deep neural network as the predictor component of our proposed method. The input/output data is structured in a time-window format in which the DNN is trained to take p_f^{y}, p_f^{y+1}, ... , p_f^{y+9} in different time windows for each FoS and forecast p_f^{y+10}, p_f^{y+11}, p_f^{y+12}, p_f^{y+13}, and p_f^{y+14} in the output. Among all the introduced problems solved applying DNNs, some of them are fundamentally sequential in nature, meaning that the values are consecutive and the order in the data sequence is of importance. Speech recognition and machine translation problems are known as outstanding examples in this research works' domain (Deng et al., 2013). One of the best-performing methods in these sequential problems is Recurrent Neural Networks (RNN) (Mandic and Chambers, 2001) as well as its subdivisions such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Chung et al., 2014).

RNNs use backpropagation through time (BPTT) for gradient and loss computations and tuning the networks' weights accordingly. It could result in vanishing or exploding of various inputs since the outputs will be very small or very large quantities (Hochreiter and Schmidhuber, 1997; Goodfellow et al., 2016). Long Short-Term Memory RNNs are designed to overcome this issue using a memory unit (*aka* memory cell) to memorize the recently calculated values in data vectors. In LSTM, the network gains more control over the recurring weight multiplications and is able to stop the network from producing ineffectual results in the output (Dargan et al., 2019). This mechanism also allows the network to learn and remember the essential data and forget the insignificant parts of it; that is, the LSTM RNNs are capable of remembering the influential data by electing which information is more relevant in sequences, resulting in more accurate predictions (Chung et al., 2014; Abrishami and Aliakbary, 2019). The more these memory cells are, the more data the LSTM is able to memorize.

The state-of-the-art research indicates that LSTM architecture has resulted in significant improvements compared to the other DNN and machine learning solutions for these sequence-to-sequence predictions, such as regression problems (Chung et al., 2014; Dargan et al., 2019). The reason is that every element of the time-series could have a severe effect on the results. Therefore, they all must be given the chance to be remembered by the trained network, making this architecture the right fit for the stated problem.

Finally, we describe our proposed method's trend prediction workflow to conclude this section. Figure 1 illustrates the data pre-processing steps and the publication predictor module. In the first step, the number of published papers per year in each field of study are extracted from the dataset to create a database of all research areas in computer science studies. This database contains the number of publications per FoS in a 78 years time span, starting from 1940 to 2017. In step 2, all level-one fields of study in the database are separated and stored as a vector. Level-one fields of study consist of the most general FoSs in computer science research. In step 3, time windows are generated so that each one consists of all level-one FoSs data in a fixed window length of 15 years. These overlapping windows start in the year 1940, adding 1 year to each starting point until the final window is generated. This step augments the dataset to enhance the deep learning architectures' performance. Step 4 integrates all these time windows into a single dataset, while step 5 separates the input data from the expected results. In step 5, the training and test sets are also separated so that the test set contains the ground truth values of years 2013 to 2017, which are never observed in the training set. It is essential to note that the test set is selected to resemble the actual prediction situation; therefore, it is not shuffled or randomly selected. Ultimately, in step 6, the LSTM neural network is trained with 10 years of input data and the corresponding 5 years of the expected results. In the training phase, the data from 1940 to 2012 is fed to the network in two randomly separated sets of training and validation with an 80/20 ratio. This validation set helps the training network to avoid underfitting and overfitting to the training data. Then the trained architecture's performance is evaluated with the test set. Further predictions for the following years not included in the dataset are also performed by employing 2008 to 2017 data as the input for each study field. These steps are also redone for specific level-2 fields of study to further analyze the prediction results from level 1.

Implementation details

In this section, the implementation details and the structure of the proposed LSTM model are introduced. Tensorflow's *Keras* library is used to build a sequential model. The input layer accepts ten different time-series values of each FoS, followed by an

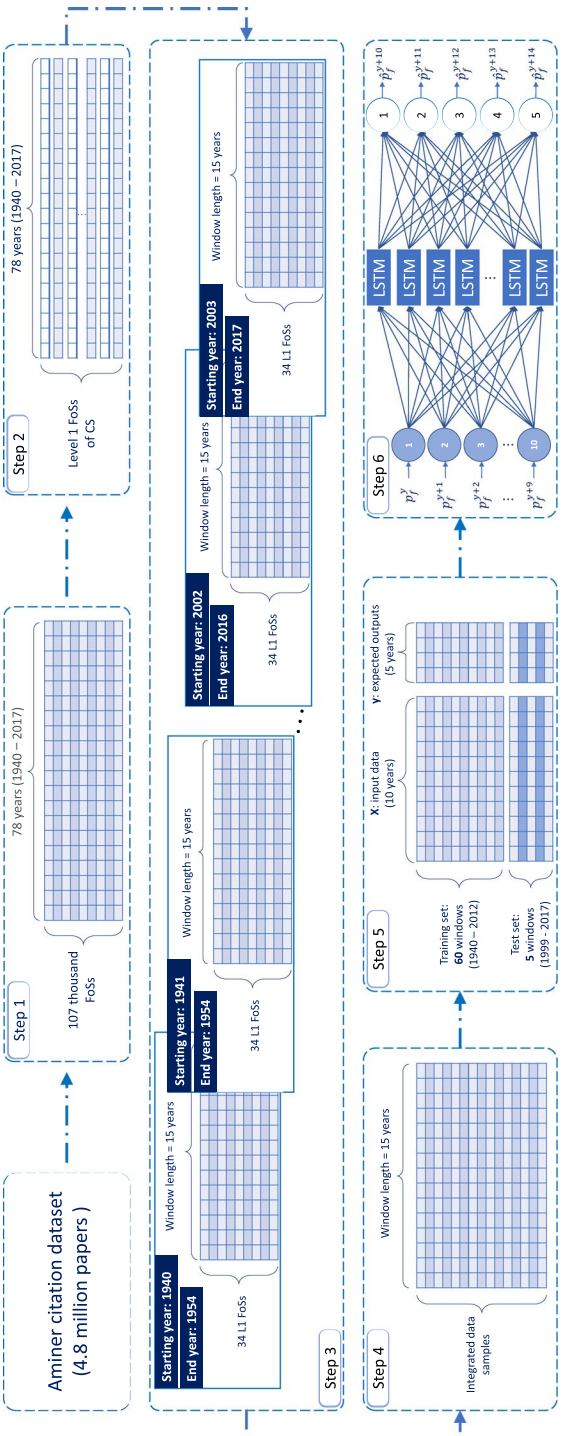


Fig. 1 A schematic of the proposed method's workflow

Table 2 Summary of the implementation details of the proposed method

Neural network API	Keras
RNN module	LSTM (units = 100, dropout = 0.1)
Activation function	Rectified Linear Unit (ReLU)
Output dimensions	5
Output layer type	Dense
Overfitting prevention technique	Early stopping (patience = 500)
Epochs	2000
Optimization algorithm	RMSProp
Learning rate	10^{-3}

LSTM layer, utilized to learn the consecutive structure of the data and a 5-dimensional output layer presents the prediction outcome $\hat{p}_i$. Input and output layers are built using Keras *'Dense'* layers, and the LSTM employs a *Rectified Linear Unit (ReLU)* activation function formulated as $f(x) = \max(0, x)$. This layer includes 100 internal units, and a dropout of 0.1 is established on this layer to regularize its training process. The implemented model is optimized using the *RMSProp* optimization algorithm with a learning rate of 0.001, and the loss of each epoch is calculated with the RMSE function. A validation split of 0.2 is also set to the fit function in order to inspect the training process's quality. The coefficient of determination is also defined as a metric to be calculated throughout this phase. The learning process is executed in 2000 epochs, and an early stopping with a *patience* of 500 is applied to the validation set's loss in order to prevent overfitting to the training dataset. The remainders of the parameters and hyperparameters are set to the *Keras* default values.

The model was configured in various steps using different architectures with a single LSTM layer, stacked LSTMs, and combinations of LSTM and fully connected layers. Each architecture was trained and tested multiple times using different sets of hyperparameters to verify that the results were not randomly generated. The hyperparameter tuning was manually performed based on the authors' experience and multiple trials and errors. Comparison results demonstrated that a single LSTM layer with a relatively large number of units and a small amount of dropout could outperform deeper architectures. Different values have been examined for the rest of this architecture's hyperparameters, and the model went through the training and evaluation process ten times with each set of parameters. The final evaluation was conducted by calculating the metrics' mean to avoid random outcomes, and the best combination of the hyperparameters was chosen as the proposed method's predictor module. Table 2 encompasses the final implementation details.

Evaluations

This section describes the evaluation process of the proposed method. After introducing the adapted dataset, the baseline prediction methods, and the proposed evaluation metrics, the assessment results will be discussed. Subsequently, The suggested trend prediction method will be compared against baseline approaches.

Dataset

The DBLP Citation Network dataset[2] provided by Tang et al. (2008) is employed for this research. This dataset is provided by aggregating the Microsoft Academic Graph (MAG)[3] (Sinha et al., 2015) and the DBLP[4] dataset on computer science articles. The 12th version of the dataset used in this work consists of nearly 4.9 million papers and 45 million citation relationships and was updated in April 2020. Although most of the recent papers are available in this dataset, it is observed that the data is not reliable for the last three years due to incompleteness. As a result, we consider 2017 as the last year with reliable data for our purpose.

Each research paper is addressed in a JSON file consisting of the paper's ID, title, set of the authors and their corresponding affiliations, the publisher venue, the year of publication, the cumulative number of citations, and a set of references specified by their IDs. Other supplementary fields such as the abstract, language, DOI, and other URLs might be available for some papers. Several Fields of Study (FoS), introduced in the following, are also available for all papers.

The level-one subdomains of the Computer Science studies were extracted from the Microsoft Academic's website documents. The database was then queried to count the number of published papers labeled with each level-one CS subdomain, which resulted in a dataset of the paper count of each FoS in each year. This dataset then went through the preprocessing and data augmentation process as explained in "Publication trend prediction" section, in which 15-step time windows were generated and separated with a 10 to 5 ratio to define the inputs and outputs of the model.

Field of study (FoS)

The topic distinction in this scholarly work is based on the FoS field provided by the DBLP citation network. Each paper in this dataset is given a set of FoSs and their corresponding weights. These fields of study are directly taken from the Microsoft Academic Graph, and are arranged in a four-level hierarchical format. This hierarchy is adapted from Shen et al.'s (2018) method to build a scientific concept ontology by crawling Wikipedia articles on a weekly basis as the source of concept discovery. Their work is then followed by a multi-label classification to assign each paper with related concepts according to their titles and abstracts. The word vectors of paper abstracts are compared against the fields of study using cosine similarity, and a threshold of 0.5 (in a scale of [0,1]) is set to omit the less related topics to each paper. Finally, a concept hierarchy is built as a Directed Acyclic Graph (DAG) adapting the Sanderson and Croft's model (Sanderson and Croft, 1999). It is assumed that if concept x is seen in 80% of all the papers, including concept y, then concept x is a subset of y. For example, more than 80% of papers containing the *"Data Mining"* label also include *"Computer Science"*, while the opposite is not valid. Therefore, *"Data Mining"* is a subset of *"Computer Science"*.

Another remarkable point is that the concept hierarchy built by Shen's method is a DAG (Shen et al., 2018); therefore, a node can have multiple parents as well as multiple

[2] aminer.org/citation.

[3] microsoft.com/en-us/research/project/microsoft-academic-graph.

[4] dblp.org.

Fig. 2 World cloud of current CS topics on the first-level fields of studies. The larger text size demonstrates a higher number of published papers in the corresponding research area

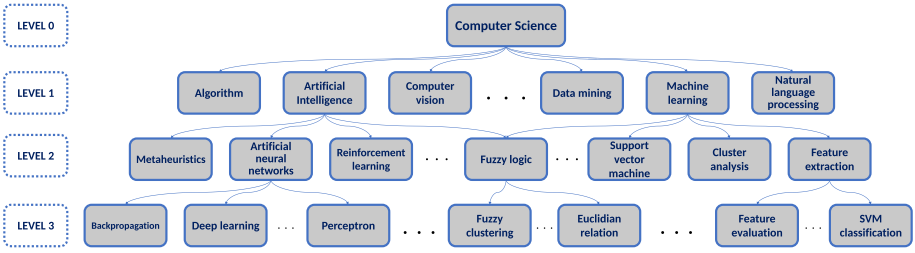


Fig. 3 A subsection of Computer Science topics' hierarchical structure and its subsets in the MAG dataset

children. As an example, *Artificial Neural Network* is a child of both *Artificial Intelligence* and *Machine Learning*. This technique is applied to construct a four-level hierarchy in the citation network dataset.

Level zero of the topics contains the main scientific areas such as art, mathematics, chemistry, and computer science. On the first level of Computer Science subsets, 34 of the primary CS research areas exist; including *Algorithms*, *Artificial Intelligence*, *Data Mining*, *World Wide Web*, etc. Current level-one CS research areas are shown in Fig. 2, where the text size correlates to the number of published papers. The primary goal of this research is to predict the trending areas on these 34 principal subsets of the CS studies.

Moreover, all these research areas encompass tens of more branches in the lower levels two and three, becoming more specific as we move downwards. A section of this hierarchical topics graph is shown in Fig. 3.

Baseline methods

Three different baseline methods are taken into consideration for the sake of comparison in this section. One naive prediction method, one SVM-based learning method, and the Hurtado et al's (2016) method are going to be compared with our proposed trend prediction approach. These approaches are introduced in the following:

Last Year is a naive prediction method that estimates the future number of publications by presenting the last valid year's data in hand. For instance, if 2017 is the last year with complete

data in our dataset, this method uses 2017's data to predict 2018's values corresponding to each FoS. It also uses the same method of prediction for any further years. Therefore, it presents the 2017's number of publications data for 2018 to 2022 in each field of study. Research has shown that last year's data presents promising results in such studies (Chen et al., 2018), although this is a simple forecasting method. Thus, outperforming this approach is not a trivial task. The formula for last year's predictions is shown in Eq. 2 where i demonstrates the last year with valid data, and the prediction is being made for the kth succeeding years' number of publications in the filed of study f.

$$\forall k \in \mathbb{N} : \hat{p}^f_{i+k} = p^f_i \tag{2}$$

Furthermore, Support vector machines (SVMs) have always been among popular methods for classification problems. Some of their advantages such as the independence of the computational complexity to the input space dimensionality, high accuracy in prediction problems, and the capability of generalization have also made them suitable for regression problems, creating the concept of support vector regression (SVR). Therefore an SVM with a linear kernel is optimized for the second baseline forecasting method.

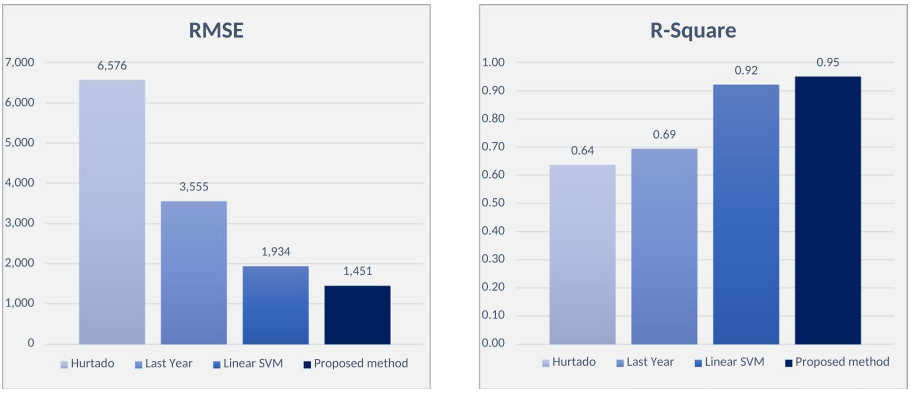
Finally, we consider Hurtado et al.'s (2016) work as the third baseline method. This paper adapts an association analysis approach in order to find the subdomains of data mining and machine learning fields of study and takes advantage of a linear-regression-based ensemble forecasting method to take different FoSs' effects on each other's trends. It is noteworthy that only the prediction method is taken into consideration for the sake of comparison since our dataset contains fields of study, and we do not need to extract the research areas from scholarly works. The paper uses nine last years' data to make these predictions on the first upcoming future year. We have expanded this approach to make 5 years of predictions so that the results are more comparative with our proposed method.

Measurement criteria

Two different metrics are taken into account for the evaluation of the proposed method and further comparison with the baseline approaches. *Root Mean Squared Error (RMSE)* measures the difference between the actual and predicted values, while the *coefficient of determination*, namely R^2 or *R-Squared*, calculates the correlation between them. R^2 is ranged between 0 to 1, where zero shows no correlation at all and one points out a perfect correspondence. RMSE takes any values equal to or larger than zero, in which zero substantiates exact errorless predictions. Therefore, larger amounts of R^2 and lower amounts of RMSE illustrate that the predicted values are more accurate and desirable. Equations 3 and 4 demonstrate RMSE and R-Squared formulas where p is a set of actual values in hand, and $\hat{p}$ encompasses the forecasted outcomes. $\overline{p}$ also indicates the mean of p_i values.

$$\mathrm{RMSE}(p, \hat{p}) = \sqrt{\frac{1}{N_{\text{samples}}} \sum_{i=1}^{N_{\text{samples}}} (p_i - \hat{p}_i)^2} \tag{3}$$

$$R^2(p, \hat{p}) = 1 - \frac{\sum_{i=1}^{N_{\text{samples}}} (p_i - \hat{p}_i)^2}{\sum_{i=1}^{N_{\text{samples}}} (p_i - \overline{p})^2} \tag{4}$$



(a) Average RMSE in 5 years. **(b)** Average R^2 score in 5 years.

Fig. 4 Overall comparison of 5 years' prediction results between the proposed method and the baselines

Evaluation results

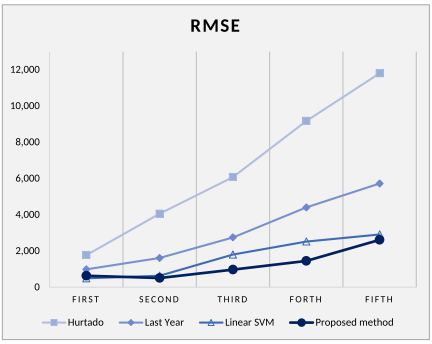
This section is dedicated to the evaluation of our proposed method in different scenarios. After the model is trained using the training set, its predictions on the test set to forecast data from 2013 to 2017 are going to be compared with the ground truth data of the same period, and the performance will be assessed using the two measurement criteria introduced in the preceding section. Moreover, the baseline methods are going to make predictions for the same test set, and the results will be compared against the proposed method.

It is noteworthy that the proposed method uses the past 10 years for making predictions in all the following experiments. However, the Last Year, Linear SVM, and Hurtado et al.'s approaches use 1, 9, and 10 years of recent data, respectively. All the predictions are made on the level-one data of the CS publications.

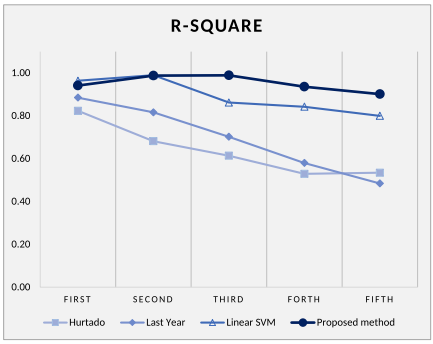
In the first part of the assessments, the overall forecasting accuracy of the proposed method in 5 years ($\hat{P}$) is measured against the baselines. Figure 4 illustrates a comparison of the overall RMSE and R-squared of these predictions on different methods. It is shown that the proposed method outperforms the state-of-the-art prediction accuracy in terms of both accumulated RMSE and R-Squared of the entire five forecasted years.

Next, the prediction performances of different schemas are observed in a yearly manner. Therefore, we can decide which method is more reliable to forecast each upcoming year in the future. The aim of this part of the evaluation is to find the most distinctive approach in predicting p_f^{y+10}, p_f^{y+11}, p_f^{y+12}, p_f^{y+13}, and p_f^{y+14} separately. Figure 5 shows the comparison results of this examination.

It is observed that the proposed trend prediction approach outperforms the baseline methods' overall performance in the 5-year period regarding both measurement criteria. It also achieves comparable results in the yearly assessments and is able to outperform the baseline performance in most cases, therefore proving its robustness.



(a) Yearly RMSE in 5 years of prediction.



(b) Yearly R^2 score in 5 years of prediction.

Fig. 5 Yearly performance comparison between the proposed method and the baselines

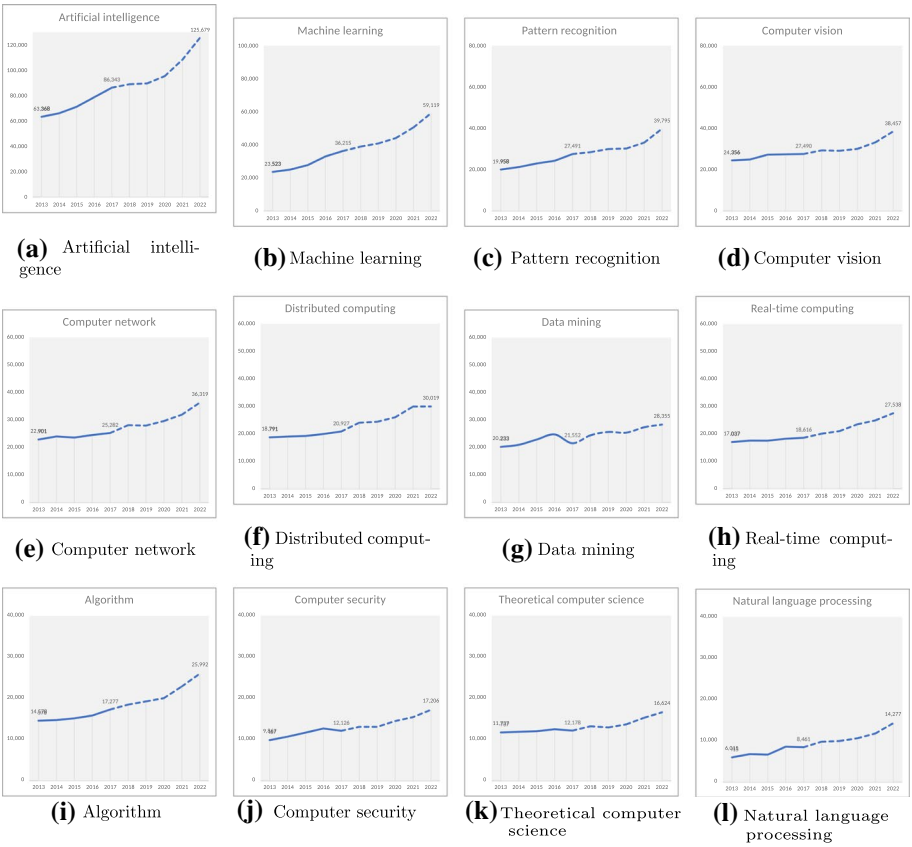
Discussion and results

" Evaluation results" section demonstrates the robustness of our proposed method. The proposed LSTM is able to outperform all the baseline methods in the overall comparison as well as the yearly assessment. After affirming our approach's accuracy, we will identify the trending areas detected by the proposed *Deep Neural Network Trend Prediction (DNTP)* in the next years to come. In this section, these predictions will be discussed in terms of various sorts of analysis.

The highest number of published papers

The primary purpose of the proposed method is to make predictions on the actual number of published papers in each research area. Therefore, the top popular fields of study in the next 5 years are going to be announced first. *Artificial Intelligence (AI)* is known to be the most popular research area of all so far, and the same condition remains in the following years as predicted by the DNTP. Followed by AI are its adjacent FoSs, namely *machine learning*, *pattern recognition*, and *computer vision*. *Computer networks* and *distributed computing* come next, perhaps due to the advent of areas such as *cloud computing*, *blockchain*, *IoT*, and *mobile networks*. Figure 6 shows the number of publications trends in the past 5 years and the five upcoming years for the top 10 trending fields of study.

Likewise, we have analyzed the subfields of *Artificial Intelligence* since it is the most popular research area by far. *Artificial neural networks* have become trending in the 2010s as a result of the advances in *deep learning* researches, and the same pattern is predicted to be preserved, resulting in the ANNs being the most favoured subfield of AI. It is followed by the scholarly works on *fuzzy logic*, *heuristics*, and *reinforcement learning* papers. Figure 7 demonstrates the top trendy FoSs in *artificial intelligence* research.



(a) Artificial intelligence **(b)** Machine learning **(c)** Pattern recognition **(d)** Computer vision

(e) Computer network **(f)** Distributed computing **(g)** Data mining **(h)** Real-time computing

(i) Algorithm **(j)** Computer security **(k)** Theoretical computer science **(l)** Natural language processing

Fig. 6 Demonstration of the number of publications' evolution for 5 years of ground truth and 5 years of predicted data concerning the top 12 CS trending areas in 2022

Top growing FoSs

Although few research works might be published in some fields of study, they can have a significant amount of growth in the next 5 years. Consequently, it is also beneficial to spot these top growing research areas for all the stakeholders in the research environment. *Computer engineering* shows a considerable amount of growth, demonstrating that the number of published papers will be more than twice in the five upcoming years, perhaps due to the engineering challenges and advancements of computer hardware design projects. It is followed by *library science*, an FoS with a relatively small number of published papers in which the considerable growth could be a result of the open science movements in these years. Figure 8 presents the top five growing areas in the 5 years to come, compared to the least growing fields of study. It is worth mentioning that each FoS's growth is only compared to its current number of publications in this assessment. The results also demonstrate that there is no computer science FoS with negative growth, and all the fields of study are going to gain more attention in the future in comparison to the present.

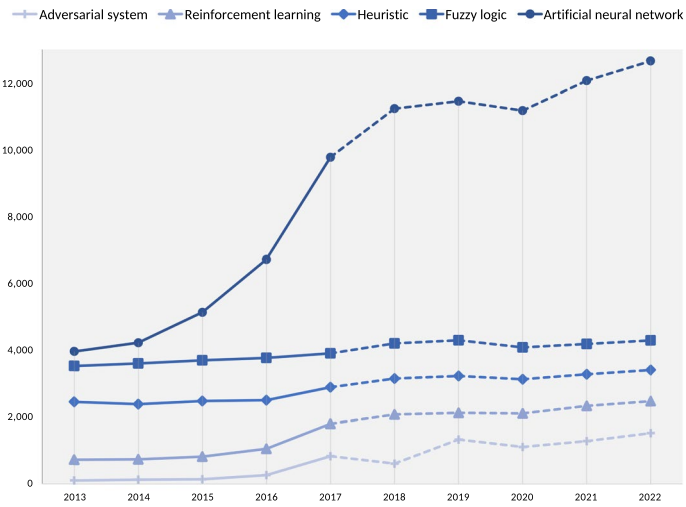


Fig. 7 The number of publications' evolution for 10 years of ground truth and predicted data concerning the top 5 trending sub-fields of *Artificial Intelligence* in 2022.

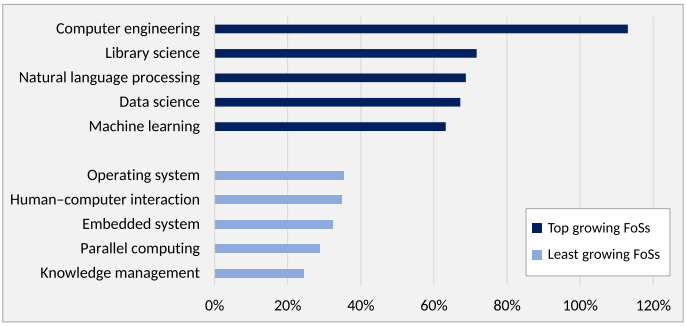


Fig. 8 Top 5 FoSs with the most growth in the upcoming 5 years (dark blue), compared to the bottom 5 of least growing areas (light blue)

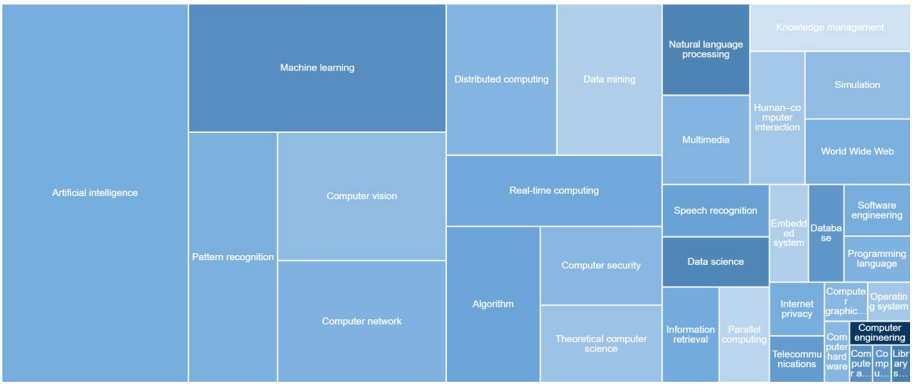


Fig. 9 Tree map of the fraction of published papers (area) in 2022 and their corresponding growth (saturation) from 2018 to 2022

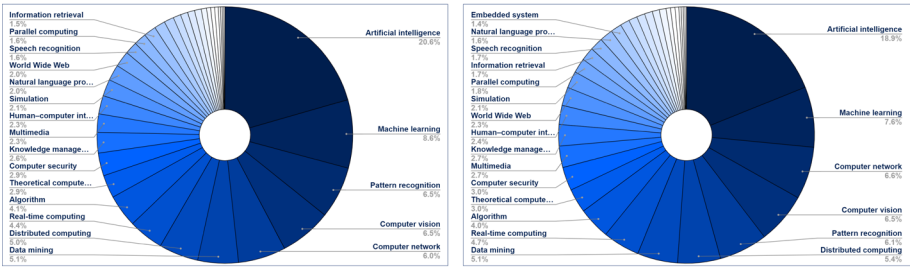


Fig. 10 Fraction of each FoS compared to the whole in the years 2017 and 2022

Table 3 A comparison of the differences between top FoS fractions in 2017 and 2022

Field of study (%)	2017 (%)	2022 (%)
Artificial intelligence	20.56	18.92
Machine learning	8.62	7.61
Computer network	6.02	6.60
Computer vision	6.55	6.45
Pattern recognition	6.55	6.11
Distributed computing	4.98	5.42
Data mining	5.13	5.11
Real-time computing	4.43	4.66
Algorithm	4.11	4.03
Theoretical computer science	2.90	3.03

Figure 9 combines the number of published papers and growth for all the CS fields of study, in which the larger area means more publications in the upcoming years and the saturation demonstrates growth. The lighter colours represent lower growth, while the darker colours show a higher rise in the number of published scholarly works in the associated research area.

Fractional analysis

Another distinctive perspective to analyze trends is the fraction of publications in each research area, comparing the division of published papers in each specified FoS between two different years. This motivation of this analysis is that only one FoS in the top 10 research areas with the highest number of publications, namely *Machine Learning*, appears in the most growing fields. As a result, we need to discover the future distribution of fractions in the CS research areas compared to the present. Figure 10 represents how these fractions are apportioned between FoSs in the year 2017 compared to the predicted future trends. It is observed that most of the larger fractions in the present will become smaller in the upcoming years. In contrast, a higher percentage of publications will be allocated to the relatively smaller fields of study. Although *artificial intelligence*, *data science*, *computer vision*, and similar fields of study in Fig. 6 were proved to become the most trending areas in the forthcoming years, they will decline in terms of the fraction of the whole published

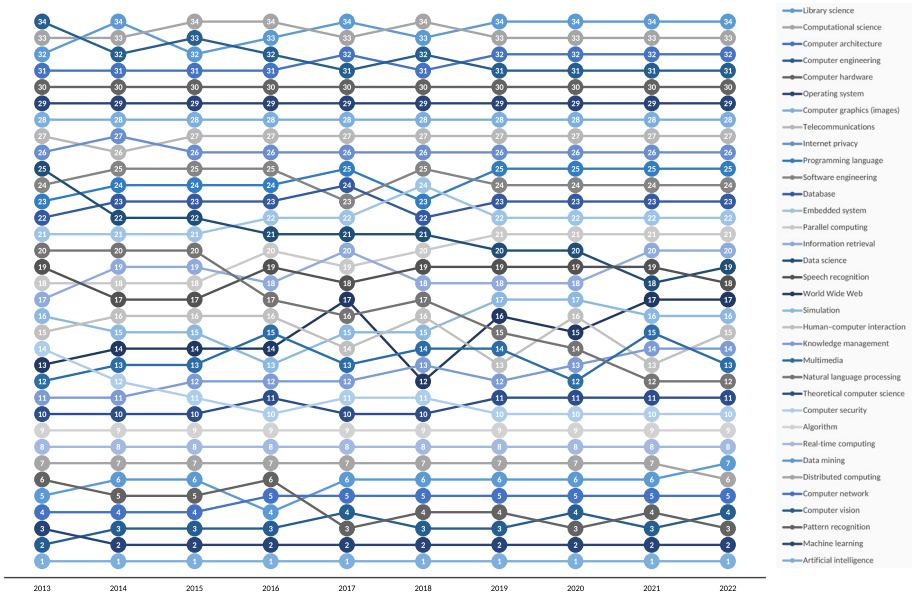


Fig. 11 Evolution of the ratings for all 34 computer science FoSs in a range of 5 past years and 5 future years of prediction

papers. Table 3 demonstrates the changes of top fields of study between the actual data of 2017 and the predicted results of 2022.

Comparing the shifts between the two charts in figure 10 indicates that the more fundamental FoSs with lengthier backgrounds such as *computer networks*, *multimedia*, and the *world wide web* will evolve and take more parts in the CS research to answer the latest problems in the research society and the industry. It also determines that the above-mentioned fields of study are becoming closer to the more trending research areas such as *artificial intelligence*, *machine learning*, and *data mining*. The fact that a large amount of data is being generated on mobile devices more than ever, and they are more accessible via the web, creates an excellent opportunity for the data-driven and AI methods to take advantage of them. It results in more interdisciplinary scholarly works between the mentioned fields of study, namely *computer networks*, *multimedia*, and the *world wide web* and AI-based methods; which brings about a significant growth in the fraction of these researches.

Evolution of ranks

This section analyzes how different FoSs of computer science have evolved in the past and will evolve in the years to come concerning their rankings. Furthermore, we will investigate if the growth in their associated number of published papers and the fraction of all research papers each FoS is occupying can result in a change of their corresponding ranking or not. *Artificial intelligence* and *machine learning* are the top two ranks with no change in the previous years, nor will they change in the near future. The FoS which has experienced the biggest change in the ranks below five in the studied period is in *pattern recognition*; which started from rank 6 in 2013 and is predicted to be among the top three

trending areas. The FoS with the most dramatic change in this period is *natural language processing* which started from rank 20, and is predicted to end up at 12 in the year 2022, perhaps due to the advancements of deep learning NLP methods. *Computer security* also experiences a four-step growth in the ranking as a result of research works on protecting cloud-based and IoT-based technologies as well as blockchain research. Figure 11 illustrates the evolution of rankings for each CS field of study from 2013 to 2017, followed by the predictions for the next 5 years.

Conclusion

Computer Science is among the most evolving domains, and therefore, analyzing its research direction can be valuable to many stakeholders. This paper proposed a novel approach for predicting the number of published papers and the top trending research areas in computer science research to predict its future directions. A deep neural network model is applied for learning the above-mentioned forecasting task based on a sequential time series of 10 past years of the number of published papers in each field of study. An extensive evaluation proves that the proposed method can outperform the state-of-the-art approaches in terms of overall and annual accuracy. Moreover, the forecasted trending areas were discussed from various perspectives. These investigations can provide a better outlook from the future of CS studies to the active research community, and the top trending areas can be considered for more successful budgeting. The same trend prediction approach can be applied to lower levels in the hierarchy of concepts to define optimal career paths in each principal field of study.

As future works of this research, more studies can be done on the relationships and the distance between these fields of study and how they affect one another to result in emerging trends. Further analysis can also be done on the effects of citation trends and how they influence publication trends and vice versa. We will also consider applying the current method to CS-related fields of study in order to explore the trending interdisciplinary works due to the advances in computer science research.

References

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

Behrouzi, S., Sarmoor, Z. S., Hajsadeghi, K., & Kavousi, K. (2020). Predicting scientific research trends based on link prediction in keyword networks. *Journal of Informetrics, 14*(4), 101079.

Brockwell, P. J., & Davis, R. A. (2013). Stationary ara processes, time series. *Theory and Methods, 3,* 77–110.

Chen, C., Wang, Z., Li, W., & Sun, X. (2018) Modeling scientific influence for research trending topic prediction. In *AAAI* (pp. 2111–2118).

Cheng, Q., Xin, L., Liu, Z., & Huang, J. (2015). Mining research trends with anomaly detection models: The case of social computing research. *Scientometrics, 103*(2), 453–469.

Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications, 83,* 187–205.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Clauset, A., Larremore, D. B., & Sinatra, R. (2017). Data-driven predictions in the science of science. *Science, 355*(6324), 477–480.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*(6), 066111.

Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2019). A survey of deep learning and its applications: A new paradigm to machine learning. *Archives of Computational Methods in Engineering, 27,* 1–22.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association, 69*(345), 118–121.

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8599–8603). IEEE.

Dridi, A., Gaber, M. M., Azad, R. M. A., & Bhogal, J. (2020). Scholarly data mining: A systematic review of its applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11,* e1395.

Ebadi, A., Tremblay, S., Goutte, C., & Schiffauerova, A. (2020). Application of machine learning techniques to assess the trends and alignment of the funded research output. *Journal of Informetrics, 14*(2), 101018.

Effendy, S., & Yap, R. H. C. (2017) Analysing trends in computer science research: A preliminary study using the microsoft academic graph. In *Proceedings of the 26th international conference on world wide web companion* (pp. 1245–1250). International World Wide Web Conferences Steering Committee.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., et al. (2018). Science of science. *Science, 359,* 6379.

Garousi, V., & Ruhe, G. (2013). A bibliometric/geographic assessment of 40 years of software engineering research (1969–2009). *International Journal of Software Engineering and Knowledge Engineering, 23*(09), 1343–1366.

Goodall, A. H. (2006). Should top universities be led by top researchers and are they? A citations analysis. *Journal of Documentation, 62*(3), 388–411.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). MIT Press.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.

Hegselmann, R., Krause, U., et al. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation, 5,* 3.

Hochreiter, S., & Schmidhuber, J. (1997) Lstm can solve hard long time lag problems. In *Advances in neural information processing systems* (pp. 473–479).

Hoonlor, A., Szymanski, B. K., & Zaki, M. J. (2013). Trends in computer science research. *Communications of the ACM, 56*(10), 74–83.

Hurtado, J. L., Agarwal, A., & Zhu, X. (2016). Topic discovery and future trend forecasting for texts. *Journal of Big Data, 3*(1), 7.

Jabłońska-Sabuka, M., Sitarz, R., & Kraslawski, A. (2014). Forecasting research trends using population dynamics model with burgers' type interaction. *Journal of Informetrics, 8*(1), 111–122.

Katsurai, M., & Ono, S. (2019). Trendnets: Mapping emerging research trends from dynamic co-word networks via sparse representation. *Scientometrics, 121*(3), 1583–1598.

Krenn, M., & Zeilinger, A. (2020). Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences, 117*(4), 1910–1916.

Leydesdorff, L. (2001). *The challenge of scientometrics: The development, measurement, and self-organization of scientific communications*. Universal-Publishers.

Mahalakshmi, G. S., Selvi, G. M., & Sendhilkumar, S. (2017) A bibliometric analysis of journal of informetrics—A decade study. In *2017 Second international conference on recent trends and challenges in computational models (ICRTCCM)* (pp. 222–227). IEEE.

Mandic, D., & Chambers, J. (2001). *Recurrent neural networks for prediction: Learning algorithms, architectures and stability*. Wiley.

Pham, M. C., Klamma, R., & Jarke, M. (2011). Development of computer science disciplines: A social network analysis approach. *Social Network Analysis and Mining, 1*(4), 321–340.

Poznanski, A., & Wolf, L. (2016) Cnn-n-gram for handwriting word recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2305–2314).

Rzhetsky, A., Foster, J. G., Foster, I. T., & Evans, J. A. (2015). Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences, 112*(47), 14569–14574.

Salatino, A. A., Osborne, F., & Motta, E. (2018) Augur: Forecasting the emergence of new research topics. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries* (pp. 303–312).

Sanderson, M., & Croft, B. (1999) Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 206–213).

Sari, N., Widodo, A., et al. (2012). Trend prediction for computer science research topics using extreme learning machine. *Procedia Engineering, 50,* 871–881.

Shen, Z., Ma, H., & Wang, K. (2018). A web-scale system for scientific knowledge exploration. arXiv preprint arXiv:1805.12216.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B. J., & Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web* (pp. 243–246).

Sitarz, R., & Kraslawski, A. (2012) Application of semantic and lexical analysis to technology forecasting by trend analysis-thematic clusters in separation processes. In *Computer aided chemical engineering* (Vol. 30, pp. 437–441). Elsevier.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 990–998).

Tom, Y., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine, 13*(3), 55–75.

Tseng, Y. H., Lin, Y. I., Lee, Y. Y., Hung, W. C., & Lee, C. H. (2009). A comparison of methods for detecting hot topics. *Scientometrics, 81*(1), 73–90.

Wang, L., & Sng, D. (2015). Deep learning algorithms with applications to video analytics for a smart city: A survey. arXiv preprint arXiv:1512.03131.

Wang, Z., Li, B., & Ma, Y. (2014) An analysis of research in software engineering: Assessment and trends. arXiv preprint arXiv:1407.4903.

Wu, Y., Venkatramanan, S., & Chiu, D. M. (2016). Research collaboration and topic trends in computer science based on top active authors. *PeerJ Computer Science, 2,* e41.

Xia, F., Wang, W., Bekele, T. M., & Liu, H. (2017). Big scholarly data: A survey. *IEEE Transactions on Big Data, 3*(1), 18–35.