

Time Sequence Based Dynamic Hirsch Index Measure for Scholar Impact Factor

Qing Huo^{1,2†}, Yanwen Li^{2†}, Yue Zhao^{1(✉)}, Sijia Ma¹, Wenhua Ming³ and Zhiyong Li⁴

¹ School of Information Engineering, Minzu University of China, Beijing 100081, China
zhaoyueso@muc.edu.cn

² Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China

³ College of Traditional Chinese Medicine, Shandong University of Traditional Chinese Medicine, Jinan 250355, China

⁴ Institute of Chinese Material Medical, China Academy of Chinese Medical Sciences, Beijing 100700, China

[†] Qing Huo and Yanwen Li contributed equally to this work and should be considered co-first authors

Abstract. The Hirsch Index (H-index) has become a widely used index for evaluating the academic influence of scholars, which is of great significance for talent evaluation and resource allocation. Currently, the measure of the H-index is mainly based on some static and present-day academic features of scholars using different combinations of features. To gain a more comprehensive understanding of scholars' academic trajectories, it is necessary to consider their historical H-index for the future scholar impact factor. In this paper, we introduce the time sequence into the H-index and analyze the dynamic trend of the H-index of scholar's influence. Through the comparison of linear regression model, dynamic Bayesian networks (DBNs), and the sequence-to-sequence model of LSTM, the experimental results show that the LSTM model is the most effective for short-term H-index prediction. It achieves an R^2 exceeding 0.95, which surpasses the linear regression model and DBNs by 11% and 8%, respectively. Additionally, the LSTM model exhibits a significantly lower MAE of only 1.30, representing a decrease of 1.0 and 0.9 compared to the linear regression model and DBNs, respectively. But for a long-term prediction, the performance of the LSTM model becomes worse and the DBNs exhibits better performance. Our method can effectively predict the H-index on Chinese medicine scholar data and avoids the problem of feature collection compared with the traditional method based on static academic features.

Keywords: Hirsch Index, Time Sequence, LSTM Neural Network, Dynamic Bayesian Networks.

1 Introduction

Academic influence emphasizes evaluating the academic value of research output and the academic status of research subjects from the perspective of influence. It is an

important indicator for measuring the academic contribution and influence of scholars. The H-index, originally introduced by American physicist J.E. Hirsch [1], integrates both the quantity and quality of academic papers, thereby providing a comprehensive measure for assessing the academic standing of scholars.

In recent years, numerous researchers have explored the prediction of individual H-index, primarily focusing on two major directions. Firstly, scholars have explored the impact of various features on the accuracy of H-index prediction. Acuna et al. [2] utilized elastic net regularized linear regression to predict the future H-index of 3,085 neuron scientists using five key features. Their findings revealed that, compared to relying solely on the H-index, incorporating these features significantly improved prediction accuracy. McCarty et al. [3] approached the problem from the perspective of co-authorship networks, randomly selecting 238 authors and constructing an author-coauthor adjacency matrix. They discovered that features of the co-authorship network possess strong predictive power for the H-index. Dong et al. [4] shifted their attention to the field of computer science, revealing that topic authority and publication venue are crucial factors determining whether a paper contributes to the growth of an author's future H-index. Additionally, Tobias et al. [5] comprehensively considered both paper and author features, utilizing neural network methods to predict researchers' H-index at different future time points. They observed that, as time progressed, the influence of the H-index gradually waned, while the importance of other features increased, further validating the reliability of neural networks in predicting researchers' prospects. Momeni et al. [6] leveraged machine learning techniques and feature analysis to predict H-index, discovering that features based on non-prior impacts exhibited superior predictive power in the long term.

On the other hand, researchers have also explored the dynamic changes of H-index. Liu et al. [7] obtained a time series change model by analyzing the structure of different citation time series. Lv et al. [8] proposed a method to calculate historical H-index, providing insights into researchers' dynamic trends and research growth patterns. Zhang et al. [9] extracted the publication and citation data from two research teams and successfully identified different stages of scholars' scientific development through dynamic tracking and monitoring of the H-index of team members. Niu et al. [10] applied Logistic growth curve models and Gompertz curve models to fit the growth trajectories of researchers in different fields, achieving an R^2 values of approximately 0.90, demonstrating the high accuracy of the models. These studies have not only deepened our understanding of the dynamic changes of H-index but also provided powerful tools for evaluating and predicting scholars' academic achievements.

In our study, we introduce the time sequence to investigate the relationship between multiple academic features and the H-index. To achieve a more precise prediction of scholars' H-index, we employ a range of methodologies, including linear regression, DBNs and the sequence-to-sequence model of LSTM to evaluate the comprehensive academic influence for scholars. The goal is to explore more scientific and rational methods to predict the H-index.

2 Data Source and Analysis

2.1 Data Source

The study of predicting scholars' H-index holds particular significance, as it contributes to the fair and objective evaluation of scholars, fostering their recognition and enhancing the international reputation of Chinese medicine research. In this work, we collect the publication and citation data from the Year 2006 to 2023 for 200 scholars in the research field of Traditional Chinese Medicine (TCM).

Data Retrieval. The data for this study were sourced from the Web of Science Core Collection database. To ensure data accuracy and avoid confusion due to identical names, a precise search method combining "full name of the scholar" with "affiliated institution" was employed.

Data Download. Download the scholar's complete report and citation report. The complete report provides the number of papers published by each scholar and their total citation counts, while the citation reports detailed the specific citation history of each paper over multiple years.

Data Processing. Based on the data from the citation reports, we extracted a sequential list of the scholars' cumulative citation counts year by year. This step was crucial in supporting the subsequent calculation of historical H-index.

Sorting and Calculation. The cumulative citation count sequences were first sorted in descending order. Then, following the definition and calculation rules of the H-index, we computed the historical H-index for each scholar. Specifically, among the N papers published by a scholar, if there are h papers with citation counts of at least h , and the remaining $N-h$ papers have citation counts fewer than or equal to h , then the value of h is considered the scholar's H-index.

Prior research has demonstrated that, apart from traditional bibliometric indicators such as the number of published papers, citation counts, and average citations per paper, other factors including academic age [11], number of coauthors [12], and the number of referenced citations [13] also significantly impact scholars' future H-index. To improve the prediction accuracy we incorporate these factors as feature variables in our analysis. Table 1 provides a detailed list of the features used to predict scholars' future h-index.

Table 1. Academic Features for Predicting Scholars' Future H-index

Feature	Description
Publication Count	The number of academic papers published within a certain period
Citation Count	The total number of citations of academic papers
Average Citations	The average number of citations per paper
Academic Age	Starting from the publication of the first academic paper
Average Coauthors	The average number of coauthors per paper
Average References	The average number of cited references per paper
Historical H-index	Scholar's H-index of every year

2.2 Analysis of the Growth Trajectory of Scholars' H-index

Figure 1 demonstrates the growth trajectory of scholars' H-index over time. Based on the temporal changes, the scholars' H-index variations can be categorized into two major types. For the first group, the H-indexes of Scholar 4 and Scholar 5 exhibits a rapid growth pattern with significant rates. For the second group, the H-indexes of scholars demonstrates a steady growth trend, maintaining a consistent increase over the observation period. These observations align with the findings of Zhang et al. [9] on the evolving trends of H-index among outstanding talents. Scholars with rapid growth patterns demonstrate outstanding academic performance and immense potential, while those with steady growth patterns are renowned for their consistent academic contributions and stable development.

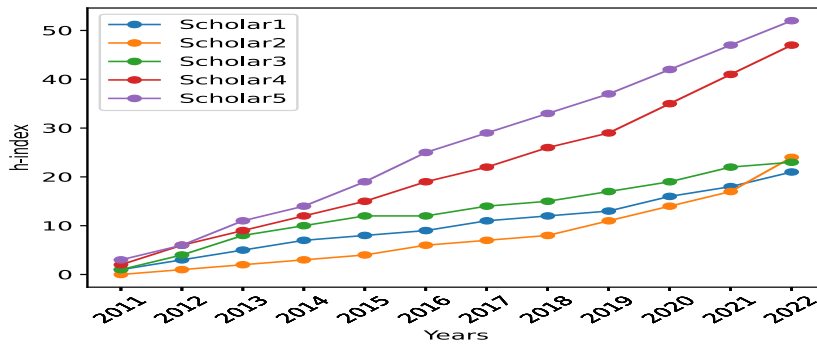


Fig. 1. Growth Curve of H-index for Selected Scholars

The rapid growth phase of a scholar's H-index can occur at various stages throughout their career, encompassing a range of intricate factors such as the quantity and quality of research outputs, academic collaborations, citation networks, and more. As an example, Scholar 2 initially lagged behind Scholar 1 and Scholar 3 in terms of their H-index during their early career. However, between 2021 and 2022, Scholar 2 experienced a surge in his H-index, surpassing both Scholar 1 and Scholar 3. This suggests that Scholar 2 may have produced innovative research outcomes, potentially indicating a golden period of accelerated H-index growth in the near future.

3 Methods

3.1 Linear Regression Prediction Model

Ayaz et al. [14] conducted a study on the impact of various features on the H-index by imposing limitations on Academic Age and H-index thresholds. They developed a linear regression model that effectively predicted the H-index of scholars in the field of computer science. In this paper, features such as Publication Count, Citation Count, Average Citations, Academic Age, Average Coauthors, and Average References were

considered as predictor variables, while the scholars' H-index served as the target variable. The model can be formally defined as follows:

$$\hat{Y} = \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \omega_4 x_4 + \omega_5 x_5 + \omega_6 x_6 + b \quad (1)$$

3.2 Dynamic Bayesian Networks

Dynamic Bayesian Networks (DBNs) consist of two components: the initial network and the transition network. The initial network, denoted as $B_0 = (G_0, \theta_0)$, G_0 represents the structure of the initial network, while θ_0 represents its parameters. It defines the probability distribution of variables at the initial time. The transition network, denoted as $B_{\rightarrow} = (G_{\rightarrow}, \theta_{\rightarrow})$, G_{\rightarrow} represents the structure of the transition network, while θ_{\rightarrow} represents its parameters. It defines the transition probabilities of variables between adjacent time slices [15]. Figure 2 shows the structure of dynamic Bayesian networks.

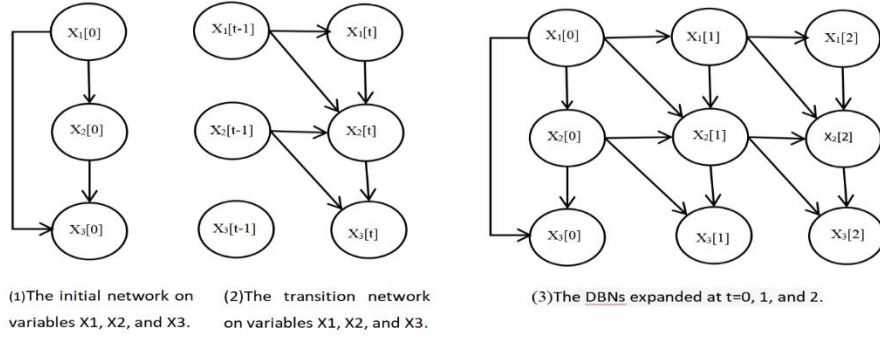


Fig. 2. Dynamic Bayesian Networks structure

The Dynamic Bayesian Networks (DBNs) defines the probability distribution over the infinitely varying trajectories within a dynamic stochastic process, ranging from the initial time point to time T . The joint probability distribution of the random process, spanning from the beginning to the end of the sequence, can be expressed as:

$$P(X_1, X_2, \dots, X_T) = P_{B_0}(X_1) \prod_{t=1}^T P_{B_{\rightarrow}}(X_t | X_{t-1}) \prod_{i=1}^N P_{B_0}(X_0^i | P_a(X_0^i)) \cdot \prod_{t=1}^T \prod_{i=1}^N P_{B_{\rightarrow}}(X_t^i | P_a(X_t^i)) \quad (2)$$

where X_0^i , X_t^i represents the value of the i -th variable at time 0 and t . $P_a(X_0^i)$, $P_a(X_t^i)$ represents its parent nodes, and N represents the number of parent nodes.

The dynamic Bayesian model regards the data as a process evolving. Its essence is to update the unknown quantity by calculating the posterior probability distribution,

given the sample state at time 1 to t . The steps to predict the H-index using a dynamic Bayesian model are:

Establish the DBNs Structure. To establish the DBNs structure, we employ a greedy search algorithm that involves the addition, deletion, and reversal of edges to construct a novel network architecture. We seek a balance between underfitting and overfitting by optimizing the network structure based on the BIC score, ultimately retaining the network structure with the optimal score. Figure 3 illustrates the dynamic Bayesian networks structure for predicting scholars' H-index.

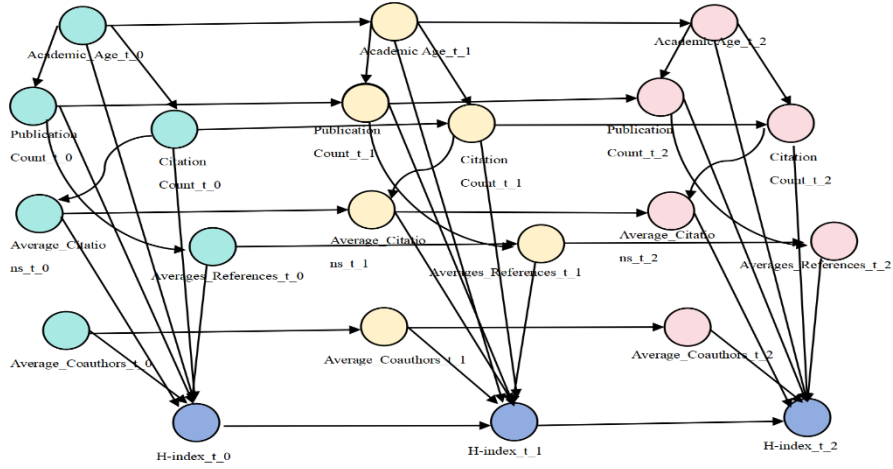


Fig. 3. Structure of the Dynamic Bayesian Networks for Scholar's H-index Prediction

DBNs Parameter Learning. Parameter learning in Bayesian networks involves the utilization of the network structure G and the training dataset D to determine the conditional probability density of each node. In this paper, we employ the maximum likelihood estimation method [16] to identify optimal parameters. The fundamental principle of maximum likelihood estimation is to find the parameters that maximize the likelihood function. The extension of this method in Bayesian networks is as follows:

$$\max \log P(D | \theta) = \sum_{i=1}^n \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \theta_{ijk} \quad (3)$$

where n represents the number of network nodes, q_i represents the total number of possible values for the parent node set, r_i represents the total number of possible values for each node, and N_{ijk} represents the total count of node i taking the k -th state when the parent node is in the j -th state.

Model Inference. This process relies on the previously established DBNs model and incorporates the explanatory variables at the time $t+1$ as the evidence for inference, achieving the prediction of the H-index at the time $t+1$. Figure 4 illustrates this inference process in detail.

Data used for modeling

				Evidence
Academic age ₀	Academic age ₁	...		Academic age _{t+1}
Publication Count ₀	Publication Count ₁	...		Publication Count _{t+1}
Citation Count ₀	Citation Count ₁	...		Citation Count _{t+1}
Average Citations ₀	Average Citations ₁	...		Average Citations _{t+1}
Average Coauthors ₀	Average Coauthors ₁	...		Average Coauthors _{t+1}
Average References ₀	Average References ₁	...		Average References _{t+1}
H-index ₀	H-index ₁	...		?

Fig. 4. The Inference Process for Predicting Scholar's H-index

3.3 LSTM

The Long Short-Term Memory Network (LSTM) [17] is a specialized type of Recurrent Neural Network (RNN) designed to handle nonlinear time-series data. LSTM effectively manages the flow of information through the introduction of intricate gating mechanisms, consisting of three crucial components: the input gate, the forget gate, and the output gate. Among these gates, the forget gate regulates the forgetting of internal state information from the previous timestep, the input gate determines the preservation of candidate state information at the current timestep, and the output gate controls the dissemination of internal state information to the external state at the current timestep. Through the synergistic action of these gating mechanisms, LSTM enables efficient processing and accurate modeling of time-series data.

Firstly, the LSTM calculates the three gates and the candidate state using the previous external state and the current input.

$$i_t = \sigma(w_i x_t + u_i h_{t-1} + b_i) \quad (4)$$

$$f_t = \sigma(w_f x_t + u_f h_{t-1} + b_f) \quad (5)$$

$$o_t = \sigma(w_o x_t + u_o h_{t-1} + b_o) \quad (6)$$

$$\tilde{c}_t = \tanh(w_c x_t + u_c h_{t-1} + b_c) \quad (7)$$

where x_t represents the current input, h_{t-1} represents the previous external input. $\sigma(\cdot)$ represents the logistic function, which outputs values in the range (0, 1), w_* represents the input weights, u_* represents the forget gate weights, and b_* represents the bias weights. Here, $* \in \{i, f, o, c\}$.

Subsequently, the LSTM update the memory cell by combining the outputs of the forget gate and the input gate, enabling effective retention and forgetting of information.

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \quad (8)$$

where \circ represents the element-wise multiplication of vectors.

Finally, the LSTM update the internal state information and transmit it to the external state, thus completing the processing and representation of the information.

$$h_t = o_t \circ \tanh(c_t) \quad (9)$$

To predict the H-index, we employed a sliding window approach for sequence construction. Specifically, we established a time window length of T and traversed the original dataset. Within each time window, the first timestep served as the starting point, and the subsequent T time steps were consecutively selected to form the input sequence. Concurrently, the future H-index values were designated as the corresponding output sequence. Through this methodology, we successfully transformed the original data into a series of input-output sequence pairs.

We employed a single-layer LSTM network with a hidden layer. The input feature dimension of the model was set to 7 to fully capture the multi-dimensional information of the data. The hidden layer comprised 256 units, providing sufficient complexity for the model to learn the underlying patterns in the data. To effectively adjust the model parameters, we selected the Adam optimizer, which is widely used in deep learning due to its adaptive learning rate adjustment strategy. Additionally, to assess the model's prediction performance, we utilized the Mean Squared Error (MSE) as the loss function, which directly reflects the deviation between the model's predictions and the actual values, thus facilitating better evaluation and optimization of the model's performance.

4 Experiments and Result Analysis

Through the comparison of the experimental results of three models on the test set, as shown as Figure 5, Figure 6 and Figure 7, we observed that the prediction results of the linear regression model are clearly inferior to both the LSTM model and the DBNs model. The DBNs encountered difficulties in predicting data at extreme value points, resulting in overly optimistic prediction outcomes. In contrast, the LSTM neural network model demonstrated the best performance among the three models.

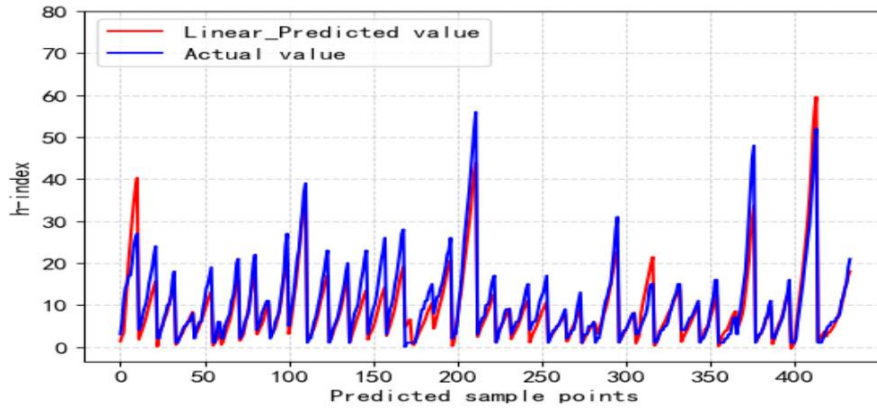


Fig. 5. Results of H-index Prediction Utilizing a Linear Regression Model

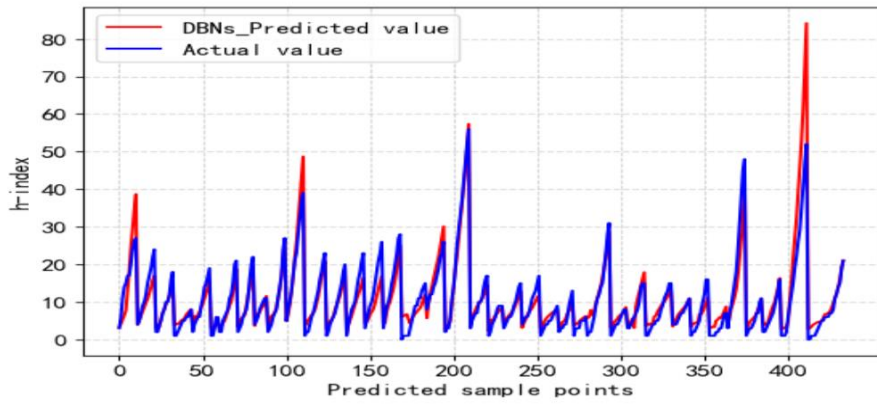


Fig. 6. Results of H-index Prediction Utilizing a Dynamic Bayesian Networks

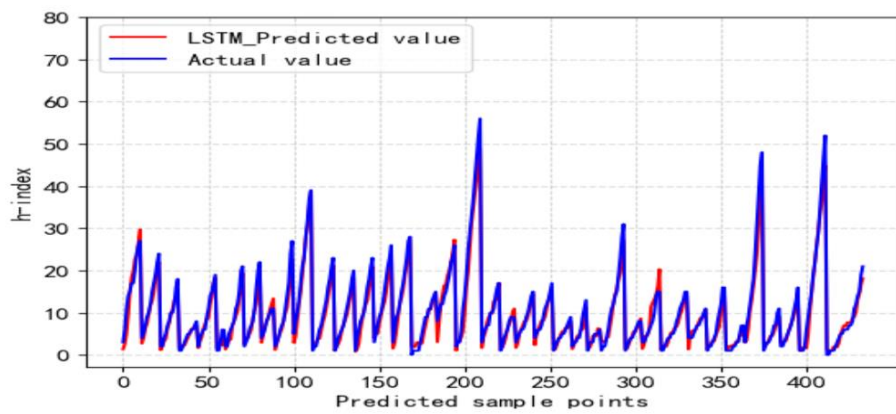


Fig. 7. Results of H-index Prediction Utilizing a LSTM Model

To further evaluate the accuracy of three models in predicting scholars' H-index, we conducted a systematic analysis of prediction errors and determination coefficients for each model, with specific data detailed in Table 2. Through comparative analysis, we found that although the linear regression model possesses a certain level of predictive power, its mean absolute error (MAE) reaches 2.34, and its determination coefficient (R^2) is only 0.84. This indicates that there may be non-negligible deviations between its prediction results and potential facts. In contrast, the DBNs significantly optimized prediction performance by incorporating advanced methods of Bayesian inference and time series modeling. Specifically, its MAE decreased to 2.23, while the determination coefficient increased to 0.87, demonstrating stronger explanatory effects and the ability to capture inherent patterns in the data. However, among the three models, the LSTM neural network model exhibited the most outstanding prediction performance. Leveraging its unique architecture and powerful learning capabilities, the LSTM model achieved an impressive MAE of only 1.30, with a determination coefficient exceeding 0.95. This firmly validates the accuracy and reliability of the LSTM neural network model in capturing the changing trends of scholars' H-index.

Table 2. Comparison of H-index Prediction Results for three Models

Model	R^2	MAE
Linear Regression	0.8424	2.3404
Dynamic Bayesian	0.8733	2.2373
LSTM	0.9535	1.3021

The prediction of scholars' H-index at various future time points not only enhances our understanding of their academic trends but also reveals their influence within their research fields. To comprehensively evaluate the predictive performance of different models, we compared the prediction results of DBNs and the LSTM model for scholars' H-index over the next n years, as shown in Figure 8. Through the comparative analysis, we can gain a more precise understanding of the dynamic changes in scholars' academic growth, providing robust support for related research and decision-making.

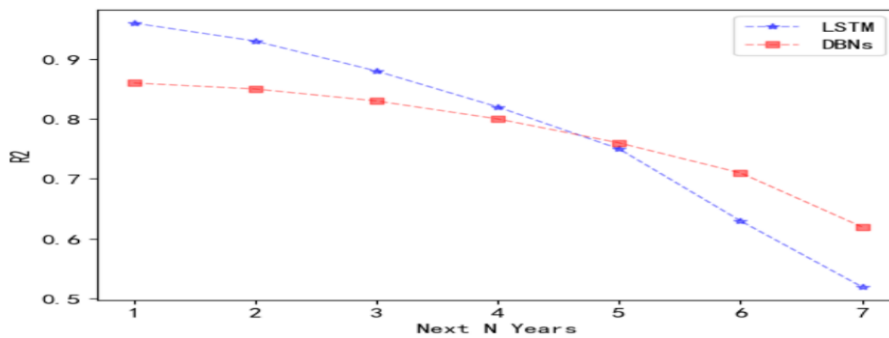


Fig. 8. Results of H-index Prediction for the Next N Years

The research findings reveal that the model demonstrates superior performance in short-term H-index prediction, exhibiting higher accuracy compared to long-term prediction. Specifically, when predicting the H-index for the next 1-4 years, the LSTM neural network model outperforms the DBNs. Within this time frame, both models achieve an R^2 value exceeding 0.8, indicating high predictive accuracy. However, as the prediction horizon extends, commencing from the fifth year onward, the R^2 value of the LSTM model experiences a significant decline, resulting in inferior predictive performance compared to DBNs. This suggests that DBNs may possess more stable performance in long-term prediction.

Due to its intricate parameter structure, the LSTM model typically requires a substantial amount of data for training and parameter adjustment. In the context of short-term prediction, the availability of numerous data points enables the model to fully leverage these data for learning and adapting to data variations, thereby achieving superior prediction results. However, when it comes to long-term prediction, the significantly reduced number of available training data poses greater challenges for the LSTM model in capturing trends within the data, potentially leading to inferior prediction performance compared to short-term prediction. Moreover, as time progresses, the noise and uncertainty factors within the data gradually increase, which may negatively impact the prediction performance of the LSTM model.

In contrast, DBNs exhibits stronger capabilities for long-term prediction by utilizing Bayesian inference to handle uncertainty issues. Bayesian inference allows the model to consider prior knowledge and uncertainty during the prediction process, thereby better adapting to the challenges of long-term prediction. This enables the DBNs to maintain good prediction performance even in scenarios with sparse data or high uncertainty.

5 Conclusion

This study aims to investigate the dynamic growth trajectory of the scholars' H-index. By collecting bibliometric data and conducting an analysis of traditional Chinese medicine data, we have observed two primary patterns of H-index growth: rapid growth and steady growth. To further reveal the underlying mechanisms behind these growth patterns, we employed time sequence modeling, successfully predicting the H-index of scholars. This research not only examines the developmental trends and evolution of scholars' H-index from a dynamic perspective, but also provides valuable reference data for talent evaluation, recruitment, and discipline development. In the future, we plan to extend this methodology to other disciplines, aiming to comprehensively reveal the growth characteristics and patterns of scholars' H-index across different fields. Additionally, we intend to integrate information from more diverse data sources, such as academic resources beyond the Web of Science Core Collection database, to apply this methodology to a broader range of disciplines.

References

1. Hirsch, J.E.: An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* 102(46), 16569–16572 (2005).
2. Acuna, D.E., Allesina, S., Kording, K.P.: Predicting scientific success. *Nature* 489(7415), 201–202 (2012).
3. McCarty, C., Jawitz, J.W., Hopkins, A., et al.: Predicting author h-index using characteristics of the co-author network. *Scientometrics* 96(2), 467–483 (2013).
4. Dong, Y., Johnson, R.A., Chawla, N.V.: Can scientific impact be predicted? *IEEE Transactions on Big Data* 2(1), 18–30 (2016).
5. Mistele, T., Price, T., Hossenfelder, S.: Predicting authors' citation counts and h-indices with a neural network. *Scientometrics* 120(1), 87–104 (2019).
6. Momeni, F., Mayr, P., Dietze, S.: Investigating the contribution of author-and publication-specific features to scholars' h-index prediction. *EPJ Data Science* 12(1), 45 (2023).
7. Liu, Y., Rousseau, R.: Definitions of time series in citation analysis with special attention to the h-index. *Journal of Informetrics* 2(3), 202–210 (2008).
8. Lv, Na.: Analysis of the H-index Dynamic Changing Trend of Scientific Research Personnel. *Information Studies: Theory & Application* 38(05), 112–115 (2015).
9. Zhang, Lin., Dong, Ying., Bi, Deqiang., et al.: H-index Trend Monitoring and Analysis of Scholars' Influence Based on Dynamic Data Integration. *Library and Information Service* 61(17), 116–121 (2017).
10. Niu, Qingao., Liang, Huimin., Zhang, Jin.: Study on the Time Series Variation of H-Index. *Journal of Information Resources Management* 10(01), 102–110 (2020).
11. Zhang, Xiaona.: Bibliometric Analysis on the Academic Age of Scholars of Library and Information Science Based on Chinese Core Journals. *Information Research* (01), 62–66 (2019).
12. Ma Rongkang, Li Zhenzhen.: High citation or zero citation: exploring the optimal scale of research cooperation based on the citation of scientific publication—evidence from the financial times TOP 45 Journals[J]. *China Soc. Sci. Technol. Inf.*, 39(11), 1182-1190 (2020).
13. Liang, Guoqiang., Hou, H., Chen, Q., et al.: Diffusion and adoption: an explanatory model of "question mark" and "rising star" articles. *Scientometrics* 124 (2020).
14. Ayaz, S., Masood, N., Islam, M. A.: Predicting scientific impact based on h-index. *Scientometrics* 114(3), 993–1010 (2018).
15. Li, X., Zhang, Y., Li, Y., et al.: Health state prediction and performance evaluation of belt conveyor based on dynamic Bayesian network in underground mining. *Shock and Vibration Mar* (16), 17-18 (2021).
16. Ji, Z., Xia, Q., Meng, G.: A review of parameter learning methods in Bayesian network. In: *Advanced Intelligent Computing Theories and Applications*, pp. 3-12. Springer, Berlin (2015).
17. Gers, F. A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with LSTM. *Neural computation* 12(10), 2451–2471 (2000).