Vol. 07, No. 3 (2025) 1877-1890, doi: 10.24874/PES07.03A.015

Proceedings on Engineering Sciences

www.pesjournal.net

AN OPTIMAL HYPER-PARAMETER TUNED TRANSFER LEARNING AIDED SCHEME FOR SCIENTIFIC PAPER CITATION RECOMMENDATION SYSTEM

Velkumar K[1]
Balasubramanian Chokkalingam

Received 10.03.2025
Revised 07.05.2025
Accepted 05.07.2025

Keywords:

Scientific Paper Citation, Recommendation System, Term Frequency-Inverse Document Frequency, Stochastic Pattern Matching, Improved Pre-Trained Inception V3 Classifier, Multi-Parameter Hunter Prey Optimization Approach

A B S T R A C T

Background: In recent times, Recommendation system (RS) shows a vibrant part in both industries and research. In academia field, there exists a need and urge to aid researchers for discovering suitable and appropriate scientific information's over recommendations. However, there is huge gap among existing RS and the real-time issues. For overcoming this existing issue, and aids research scholars to choose effective citation, a new transfer learning-based model is proposed for the citation recommendation system.

Methodology: At first, input dataset that are publicly accessible are considered and removal of stop words or functions words will be carried with the use of stochastic pattern matching approach through which the creation of library is carried. After that, the document canonicalization is performed with the use of Term Frequency-Inverse Document Frequency (TF-IDF) after which the stored data are predicted and classified by means of Improved Pre-trained Inception V3 classifier. Before classification process, the hyper-parameters are tuned on employing Multi-parameter Hunter prey optimization approach (MP-HPO). This in turn enhances the classifier performance aiding in effective recommendation of citation records. At last, ranking is done by means of employing weighted Pearson correlation coefficient to rank the predicted citation and recommend best ones by means of ranking strategy.

Result and discussion: The proposed technique designed is validated on various metrics like precision, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Recall, and F1-score. The proposed outcomes are compared with various existing schemes to show the enhancement of proposed strategy over others.

Conclusion: The validated outcome reveals the superiority of proposed model over other existing models compared.

[1] Corresponding author: Velkumar K
Email: velkumarskc@gmail.com

1. INTRODUCTION

In general, recommender system are the approaches which aids in suggesting related items for the users like products to buy, movies to watch, and so on (Liang & Lee, 2023). In past few decades, concerns like Netflix, Youtube, and Amazon have been rising continuously to meet the need for better recommendation system which enhanced exponentially for offering best recommendations as per the preference of user (Gündoğan & Kaya, 2022). This in turn predicts, the preference of user depending on the existing data. The model of recommender system was proven to be effective in making decisions, enhancing quality, and thus reducing cost of transaction in the environment of online shopping. They are satisfying huge variety of applications like education, economy, and scientific research. In past decades, the published research papers number are increasing exponentially. Because of presence of abundant data, search engines utilized earlier for the recommendation citations are not competent of offering sufficient recommendations moreover (Huynh, Dang, Nguyen, Huynh, & Nguyen, 2023; Roy & Dutta, 2022; Q. Zhang, Lu, & Jin, 2021). Typically, the citation recommendation falls under 3 categories like content-based, collaborative filtering, and hybrid recommendation system. The recommender system based on collaborative filtering predominantly process the analysis of similarity that is a paper citation matrix for the recommendations at which content dependent models considers the content and paper's meta-data like keywords and title (Y. Li, Liu, Satapathy, Wang, & Cambria, 2024).

However, the analysis of citation is the bibliographic information of the e-documents/documents that is monographs, theories, distributed catalogues, thesis/dissertation, electronic articles (Y. Huang, Wang, & Wang, 2022) and theories for comprehending the specific information for clarifying the pattern use. This in turn plays significant part in both scientific and academic research that is taking place all over the world in the knowledge disciplines. The citation recommendation might aid the researchers to identify alternative or supplementary references quickly in the massive academic resources (Y. Huang et al., 2022; Yuan, Wan, & Wang, 2023). The present research on citation recommendation focus mostly on paper citing, thus resulting in huge cited papers that are ignored which covers relations between cited papers and its citation context that were cited in citing papers (Jie, Yang, Shu, & Yanping, 2021). However, the content of cited paper was denoted often with their unique abstract and title, that were hard for acquiring and considers various citation motivations rarely.

By the advent of artificial intelligence (AI) technology, academic papers show crucial and indispensable part in academic research (Bagul & Barve, 2021). With the persistent growth in diversity and number of the academic papers, the scholars are meeting enhanced challenge to identify the literature needed in limited time. Hence, it was considered as an urgent concern in academic community: in what way to utilize and manage this huge resource of academic literature. On considering this issue, deep learning and transfer learning models (W. Huang, Liu, & Wang, 2023) are becoming emergent in this field which has several advantages than machine learning (ML) models (Chen, Shen, Pan, Tan, & Wang, 2024; Lin et al., 2025).

In recent times, DL and TL dependent schemes are highly flourishing and gaining substantial interest rate in several fields such as computer vision, and NLP (Yang, Wang, Shi, & Zhu, 2024). The influence of this models is thus pervasive indicating the effectiveness on employing information retrieved in the research recommendation system. However, with the persistent growth of unique research models, traditional frameworks designed so far are no longer suitable to meet the demand and there is a need for new framework to under research field more effectively. Henceforth, a new model is designed in this paper using hybrid transfer learning model to recommend research citations more effectively. The objective and contributions of this presented scheme are given as shown:

The main contributions of suggested model are given by:

- To remove function words and creating library with the use of stochastic pattern match approach for processing the citation record data.
- To utilize TF-IDF for performing document's canonicalization.
- To categorize data on employing Transfer Learning based classification scheme (Improved Pre-trained Inception V3) model.
- To set rank for predicted citation with the use of Weighted Pearson Correlation Coefficient model to recommend best ones.
- To assess proposed scheme performance with respect to precision, MAP, MRR, F-score, and recall.

The upcoming sections of this paper are organized as follows: Section II delivers existing review on varied recommendation models employed. The suggested scheme is presented in section III. The performance assessment is carried for proposed technique and the attained results compared with existing models are projected in section IV. To end, the proposed model is summarized in section V.

2. RELATED WORKS

Studies on several models based on citation recommendation system using machine and deep learning models are provided in this section.
The author in the work (Ko, Lee, Park, & Choi, 2022) presented a review on recent trends which links more

advanced technical concerns in the recommendation systems and were thus employed in several areas of service and those services business aspects. At first, a recommendation model for reliable analysis, data mining technique and the relevant research fields by means of application service higher than 135 articles in top-ranking list and the conferences that were published in top-tier conferences of google scholar among 2010 and 2021 were reviewed and collected. Depending on this, review was carried on recommendation system technology and models employed in the recommendation system were thus systematized and the analysis on recent trends were given.

The author in work (Urdaneta-Ponte, Mendez-Zorrilla, & Oleagordia-Ruiz, 2021) stated that in recommendation system field on education, recommended education resources relevance might enhance the process of student's learning and thus the significance on suitability and reliability make sue the relevance and helpful information. This systematic study purpose was to analyze the undertaken work on the recommendation system which supports the practice of education in the view of attaining data related to education kind and fields to be dealt with, used developmental approach and recommended elements, and thus detecting any such research gaps in future directions.

A novel citation recommendation prediction scheme was suggested in the work (A. Ma, Liu, Xu, & Dong, 2021), which was based on metadata text so as to predict the count of long-term citation, and model's core for attaining semantic information from the text of metadata. A DL based approach was employed for encoding the text of metadata, and thus extracting the higher-levels semantic features to learn the task of citation prediction. The integration of earlier citation was made to enhance the model's prediction performance.

A DL dependent model along with the well-organized dataset was presented in (Jeong, Jang, Park, & Choi, 2020) for the purpose of context aware recommendation on paper citation. This model consists of document encoder and the context encoder that employs Graph convolutional networks (GCN) layer and the Bi-directional encoder representations from transformers (BERT), which was the pre-trained scheme of text data. On modifying the relevant Peer Read dataset, a new database termed FullTextPeerRead that comprises of context sentences for citing references and meta data paper was considered. The outcomes indicate that the suggested model could achieve existing performance by giving more than 28% enhancement in terms of recall and MAP (mean average precision rate).

A context-aware citation recommendation model was presented in (Wang, Zhu, Dai, & Wang, 2020) depending on end wise memory network. This technique learns the significance of citation contexts and paper contexts representations depending on Bi-LSTM. Specifically, the citation relationship and information of author was jointly integrated in a distributed vector representation of the paper and contexts. The continuous relevance was computed among them depending on computational multilayer memory network. It also held experiments on 3 real world databases for evaluating the model performance.

A graph neural collaborative topic model which has the benefits of both relational topic models and GNN was suggested in the work (Xie, Zhu, Huang, Du, & Nie, 2021) for capturing the higher-order relationship citation and thus having high explainability due to latent semantic structure topic. The evaluation on 3 real time database reveals that the proposed model could learn better topics on comparing traditional ones.

A study was carried in (J. Zhang & Zhu, 2022) regarding the citation recommendation model from the semantic representation perspectives of the cited papers content and relations. Initially, four citation forms were extracted and designed as the relation and content of cited papers thus considering motivation of citation and co-citation relationship. Then, about 132 models were designed to generate cited papers semantic vector which covers 4 embedding models, sixteen models on integrating 4 text representation approach and 4 forms of citation contents with 112 models of fusion. At last, the similarity among the cited paper will be computed for the purpose of citation recommendation and model of quantitative evaluation depending on prediction link was designed for identifying the most suitable form of citation content and optimal models.

In (W. Li, 2024), recommendation system for scientific articles, NLP (natural language processing) and DL models were employed for processing the titles and thus extracting the articles content attributes. For this reason, initially the article titles were preprocessed and the approach of Term frequency inverse document frequency (TF-IDF) for estimating each word's importance. After that the attained attributes dimensions were reduced on employing CNN. After that, cosine similarity matrix of articles is computed depending on vectors of attributes. Likewise, the approach of predicting link was employed for analyzing the scientific articles connection in the citation network. At last, in last step, computation of two similarity matrices are integrated with the use of influence coefficient parameter for attaining final similarity matrix, and the operation of recommendation depends on higher similarity values.

A comprehensive review was made in (Gündoğan, Kaya, & Daud, 2023) regarding the recommendation system model which covers journals from two or more publisher. Here, SBERT model was employed for identifying the scope of journals similarity with that of articles. On relating the outcomes with Word2vec,

FastText, and Glove, that are mostly the preferred models in the similarity of document, it was revealed that the similarity-based sentence level recommendation using SBERT model were successful highly than others.

A model of citation recommendation was employed in (Abbas et al., 2024) with the use of DL schemes for classifying the rhetorical zones of research articles and thus computing the similarity with the use of rhetorical embedding zone which overwhelm cold-start issue. The zones of Rhetorical were the categories that were linguistically predefined thus having some usual text characteristics. The model of DL was trained with the use of ART and CORE database having 76% accuracy. The suggested scheme is applicable both in global and local context aware recommendations.

An academic paper recommendation scheme was suggested in (Wei, 2024) depending on the heterogeneous graph and attention mechanism. This model thus considers heterogenous graph and textual information for attaining the complete and rich feature representation. At last, cosine similarity was computed for generating the recommendations. The outcomes show that the model on comparing with content dependent model, recall, f value, and accuracy was enhanced. Thus, this scheme is expected to promote in-depth application of the recommendation model in the artificial intelligence field.

From the above review, it was obvious that there exist several recommendation system models for citation records using machine and deep learning, however, there exists some issues like lower performance, higher error rate and so on (Devika & Milton, 2025; S. Ma, Zhang, Zhang, & Gao, 2025). Thus, there is a need to implement an effective scheme which further enhances the performance.

3. PROPOSED WORK

The working model of the proposed citation recommendation scheme is given in detail here in Figure 1. The diagram depicting the flow of proposed system working is provided in figure 1. Primarily, the input data acquired from database considered comprises of several paper, author, and topics information which is preprocessed for removing the function words. A process of stochastic pattern matching algorithm is utilized to reduce the characters total number for enhancing the searching efficacy and similarity comparison among training and testing database is performed. After that, the document canonicalization is carried using TF-IDF approach. Finally, an improved pre-trained TL based Inception V3 scheme is employed for recommending similar or appropriate citations so as to set citation rank. At last, ranking is done by means of employing weighted Pearson correlation coefficient.

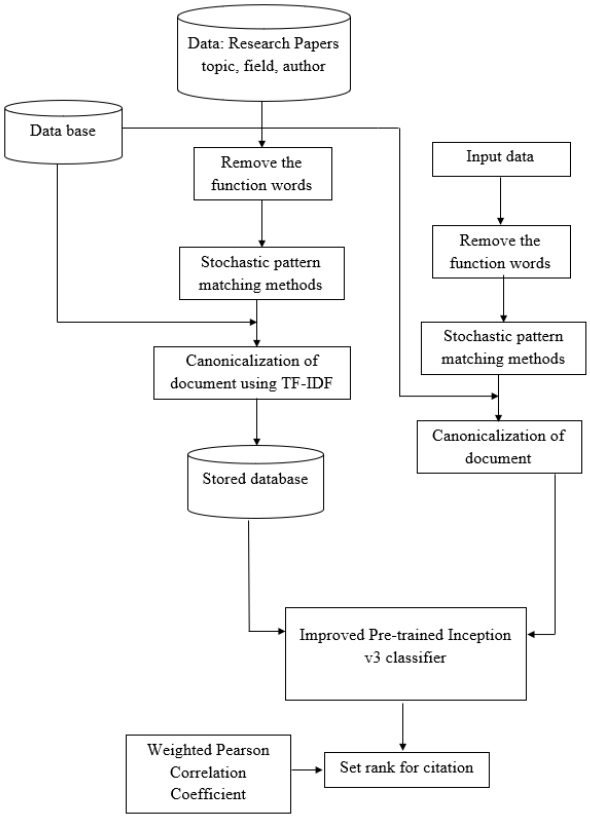

Figure 1. Flow representing working module or proposed design

3.1 Function words removal and Pre-processing input data

As an initial step, preprocessing is carried by removing function words or stop words removal to decrease the number of characters in a text. Typically, texts are the words sequence which could be segregated with the use of tokenization model. The tokenization technique is carried at which the input data are thus divided as minor parts termed tokens that eradicates the aspect of low priority. Some texts comprise of few frequent words and the ones which occurs most frequently in the given corpus will be signified as function or stop words. The function words are the one that carry less significant meaning when compared to prevalence of other words in a document. They are the most frequently used terms or words and having no semantic relations of context. The function words are the general and common words that do not exists typically contribute to semantic documents and having no added reading values. The removal of function words aids in selecting suitable words effectively thus decreasing complexity of document structure.

3.2 Library creation using Stochastic Pattern matching

Once the function words are removed, stochastic pattern matching model is used. The purpose of this is to reduce

the comparison of total number of characters between patterns and text thereby enhancing efficacy. The efficiency enhancement in searching process might be employed on using order alteration at which the character comparison is made in each attempt thus choosing shift factor that permits skipping predefined characters of text with the use of windows whose size will be equivalent to the length of pattern. This is competent of choosing whether the relation will be asserted or not thus recognizing their participants. More typical patterns are employed rather than syntactic patterns like activities, events, states, and its relations which is due to the fact that there is no need for expressing casual participant relation as complete clause. This approach is employed for the process of matching for low to high appearance match probability letters in natural English pattern string, if any mismatch occurrence, then the function Badchar () will be employed for realizing the pointer's host string forward jump thereby carrying matching priority for low probability characters based on the priority. By this, various occurrences of mismatch are predicted and several comparison times are reduced as much as possible so that the function of Badchar () will be employed to realize the string host of forward jump pointer. Therefore, by means of this, the creation of library is carried and the pseudocode representing this is shown below:

Algorithm 1: Creation of library using stochastic pattern matching model

Input: Samples of data for existing paper D_e, pattern matching database PM_d

Output: Creation of library C_L

For $i = D_e//Number\ of\ paper\ to\ compare$

  For $K = D_e(i)$

    $F_w = D_e(k), split$

  End "K" loop

  $For\ j = F_w \quad //Number\ of\ words\ in\ paper$

    $P\ (F_w) = length\ (M_w)/\text{length}\ (\text{P}M_d)$

  $SP\ (j) = Max\ (P(F_w))$

  End "i" loop

$TL\ (i)\ =\ SP$

$End\ "i"\ loop$

By means of employing this stochastic pattern matching model, creation of library is created. The document canonicalization is explained in following segment.

3.3 TF-IDF for document canonicalization

For the purpose of document canonicalization, extraction of features is carried by means of TF-IDF (Term frequency-inverse document frequency) which is employed for the extraction of weighted features from text data. It offers weight of each term at the corpus to enhance learning models. It is the product of Term frequency (TF) and Inverse document frequency (IDF). The computation of TF is shown as follows:

$$TF(t,d) = \frac{n_t}{N_{(T,d)}} \tag{1}$$

Here, n_t denotes the term t occurrences number in document d, however, $N_{(T,d)}$ denotes the document's total term T. The IDF term denotes the way at which it is important at entire corpus and will be computed using the following equation:

$$IDF = log\frac{D}{n_d} \tag{2}$$

In this, 'D' is the total corpus number in which document number is denoted by n_d where there is an occurrence of t. On employing this TF and IDF expression, the computation of TF-IDF is carried by means of following notation:

$$TF - IDF = TF * IDF \tag{3}$$

The usual procedure denoting this process is given below:

- Step 1: white space normalization: Here, the additional whitespace characters such as multiple spaces, line breaks, and tabs will be removed.
- Step 2: In this, the consecutive whitespace is thus replaced by means of single space.
- Step 3: Conversion of lowercase: The entire texts are thus transformed for eliminating the sensitivity cases.
- Step 4: Punctuation of removal: At this step, marks of punctuation are deleted such as periods, commas, exclamation marks, and question marks. Moreover, some marks of punctuation if they could bear important data like apostrophes in contractions.
- Step 5: Special characters handling: Replace or remove the special characters that do not contributes the document meaning such as currency symbols or diacritics. In the similar time, special character handling is considered based on the specified need for document or application.
- Step 6: Abbreviation expansion: usual abbreviations will be replaced by full forms for ensuring consistency and clarity. On behalf of this, lookup tables or dictionary are created for commonly employed abbreviations or expansions.
- Step 7: Synonyms resolving: Identify and replace phrases of synonyms or the words with that of single canonical process. By this, consider predefined usage of thesaurus and mappings for synonyms standardization.

- Step 8: Stop words eradication: Typical function words or the stop words will be eradicated that do not pose any such significant meaning (for example: "and", "is", "the"). The predefined stop words list that offers such kind of function is employed.
- Step 9: Stemming or Lemmatization: The words are shortened for their base form or root to consolidate similar terms. It is employed to attain this for instance lemmatization, porter stemming, and WordNet.
- Step 10: Numerical values removal: Numerical values are eliminated or replace them using placeholder token, in specific the desired numbers are not that critical one for processing or analyzing.
- Step 11: De-duplication: Find or eradicate the near duplicate or the contents of duplicate to ensure single canonical form representation. Employ approaches such as similarity measures or the function of hashing to eliminate or identify duplicates.
- Step 12: Completion and keeping canonicalized documents for the further processing analysis.

Henceforth, the issues of misspelling will be resolved for keywords used in information extraction process.

3.4 Hyperparameter tuning by Multi-parameter Hunter prey optimization approach (MP-HPO)

The hyperparameter tuning is essential before classification or prediction process so as to attain best fitness function values. This in turn enhances the classifier performance. For this purpose, Multi-parameter Hunter prey optimization (MP-HPO) model so as to tune the parameters of transfer learning employed. This approach is a recent optimization model employed to resolve the optimization issue. The technique is simulated with the use of hunting approach of some carnivorous like leopard, tiger and lions along with prey including antelopes, and deer. Here, MP-HPO is employed for solving various optimization issues.

It is regarded as advanced nature-inspired optimization model designed for enhancing the classification model accuracy. This in turn mimics interaction among prey and hunters by leveraging the dynamic balance among exploration and exploitation. It enhances the accuracy on comparing other traditional models because of its ability to optimize multiple hyperparameters simultaneously in an adaptive manner. The model introduces novel mechanism like adaptive parameter adjustment depending on predator prey proximity and hybridized model to pursuit and escape. The search efficiency is enhanced by aiding on evading local optima which is the typical limitation in other traditional models. On adjusting search patterns dynamically, this model ensures hyperparameters are tuned for maximizing predictive performance of classifier. An empirical study and benchmark experiment demonstrate that this model outperforms other such traditional models like PSO, GA, and Differential evolution (DE). The enhancement is thus attributed to their enhanced ability for adopting diverse landscape issues and their robust convergence properties. In this proposed model, hunter generally provides significance for prey distance from their group. The hunter thus modifies their position over the distance of prey whereas prey alter their location over groups. The position of search agent group might be regarded as safer location using optimum fitness function values. The process of hunter prey optimization is provided as follows:

At each iteration, each hunter with their prey location will be updated based on their rules and newer location of every member will be re-evaluated by major function. The entire member's location in initial stage might be generated randomly in the search space as follows:

$$x_j = rand(1, d) \times (ub - lb) + lb \tag{4}$$

From the above equation, number of parameter issue is denoted by d. x_i indicates the position of hunter, lb and ub resembles the minimum and maximum problem variable.

After that, the module of hunter search usually includes two stages such as exploitation and exploration. At the exploration stage, the hunter tries to search their potential region of prey. The exploitation is employed once the promising area is found, at which the hunter might minimize the behavior randomly to identify prey with their potential regions.

$$x(t+1) = x(t) + \frac{[(2CZP_{pos} - x(t)) + (2(1-C)Z\mu - x(t))]}{2} \tag{5}$$

From this equation, $x(t)$ and $x(t+1)$ denotes existing and the subsequent location of hunter, P_{pos} signifies location of prey, μ represents mean of every position, C indicates balance variable between exploration and exploitation, and their values declines from 1 to 0.02, and Z denotes the adaptive parameter:

$$C = 1 - it\left(\frac{1-0.98}{MaxIt}\right) \tag{6}$$

$$Z = \vec{R}_1 \otimes idx + \vec{R}_2 \otimes (\sim idx) \tag{7}$$

In this, it and $MaxIt$ indicates present and extreme number of iterations. $\vec{R}_1$ and $\vec{R}_2$ denotes random vector between [0,1] and number of vector index is denoted by idx that fulfills the state $\vec{R}_1 < C$. After that, based on hunting process, since the hunter was considered for prey, it dies and then hunter moves over the position of dead prey. Thus, the iteration continues with the location of prey that are being newer location of hunter. However, on assuming at which the safer position was the optimal global range position, as it offers prey having high

opportunities to survive, the hunter might select other prey by means of following expression:

$$x(t+1) = T_{pos} + CZcos(2\pi R_3) \times \left(T_{pos} - x(t)\right) \quad (8)$$

Here, previous and next location of hunter is denoted by $x(t)$ and $x(t+1)$, T_{pos} means optimal global position that covers improved fitness from primary iteration to preceding iteration, and R_3 indicates random integer.

Eventually, this is substantial in what way the prey and hunter are selected as follows:

$$x(t+1) = \begin{cases} x(t) + 0.5\left[\left(2CZP_{pos} - x(t)\right) + \left(2(1-C)Z\mu - x(t)\right)\right] \\ R_4 < \beta \\ T_{pos} + CZcos(2\pi R_3) \times (T_{pos} - x(t)) \\ R_4 \geq \beta \end{cases} \quad (9)$$

In above equation, R_4 means random integer between [0,1] and β shows parameters that are adjusting. The presented MP-HPO method attains fitness function (FF) to get better classifier performance. By this, the positive value is determined as the candidate solution of superior outcome. The presented process algorithm flowchart is specified in figure 2.

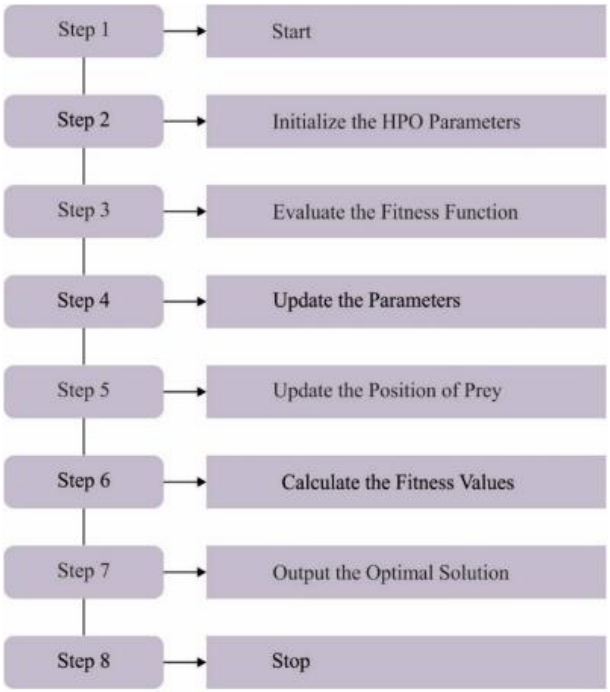

Figure 2 MP-HPO flowchart

The classifier's minimal error rate is the FF and is denoted by:

$$fitness\ (x_i) = Classifier\ Error\ rate\ (x_i) = \frac{no\ of\ misclassified\ samples}{Total\ number\ of\ samples} * 100 \quad (10)$$

Consequently, hyper parameter tuning of MP-HPO delivers best fitness function value that optimizes the module of classifier. Hence, the recommender module detects and classifies suitable citation efficiently.

3.5 Improved Pre-trained TL based Inception V3 model for classification and Weighted Pearson correlation coefficient for ranking

By the rise in depth of neural network, the neural network performance will be enhanced indeed, however this enhancement is in the expense of computing and time resources. Hence, for reducing the training cost, DL model based on transfer learning is emerging. Typically, transfer learning is employed to transfer the trained models parameters for newer model and helping in training of newer model. On considering that some tasks or data are relevant, efficiency of learning in new model could be accelerated and thus optimized on sharing parameters of trained approach to newer model over transfer learning, in spite of starting from scratch as the most models did. Due to emergence of popularity of DL, demand for data likewise increases which makes further development of transfer learning.

In this, the Inception V3 model is employed for the purpose of recommendation detection and classification shown in figure 3. This model is employed which is typically extended from the GoogleNet model. It relies on offering various small convolutions that have same level in spite of providing high convolutional ranges. The network's first layer comprises of 3 convolutions and 1 max-pooling layer. The last one comprises of channels which are non-linearly merged and provided. Hence, the reduction of network parameter is carried that implies on testing or training stage of acceleration. Similarly, extracted features enhances the classifier accuracy effectively. This scheme comprises of dense layers for enhancing the depth of network which reduces the complexity again. This model offers higher accuracy rate on comparing other existing schemes such as GoogleNet, AlexNet, and ResNet. Because of the its accuracy and flexibility, this model provides high level of performance.

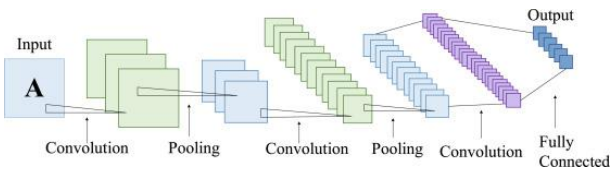

Figure 3. Inception v3 architecture model

The model presented provides ranking based functions to estimate the rank thus creating prediction or recommendation process an efficient one. The ranking value computation is carried using below equation:

$$rank = \sum_{i=1}^{M} \sum_{j=1}^{N} (A_1(i,j) + A_2(i,j))/2) \quad (11)$$

At this, A signifies the rank generated by the presented model. 'N' denotes number of classes and M is the test sets. By employing this equation, new rank that are having size $M \times N$ is accomplished. Then, on employing this, new rank and class labels are identified based on maximum value in each line (M). The pseudocode representing this scheme is shown below:

Algorithm 2: Improved Inception V3 classifier model

Input: rank1, rank2, YTest

Output: Predicted citation

Step 1: $for\ i = 1\ to\ M$

$Step$ 2: $for\ j = 1\ to\ N$

$Step$ 3: newer_score(i, j) = (score1(i, j) + score2(i, j))/2

$Step$ 4: $end\ for\ j$

$Step$ 5: $end\ for\ i$

$Step$ 6: $for\ k = 1\ to\ K$

$Step$ 7: $if\ score\ (k, 1) >= score(i, 2)$

$Step$ 8: $YPred(k) = 1;$

$Step$ 9: $else$

$Step$ 11: $YPred(k) = 1;$

$Step$ 12: $accuracy = mean\ (Ypred = YTest);$

For estimating the ranking and to set ranking for citation, the weighted Pearson correlation coefficient (W-PCC) is employed which estimates the local coefficients of each data sample points on manipulating relation metrics. Thus, this has the seizing ability to alter the relation between inputs on adding the data of spatial position which are neglected usually through previous version of PCC. The proposed PCC ρ is a dependency measure between any of two variables which are chosen randomly. The equation of this is provided by:

$$\rho = \frac{covariance(P, Q)}{\sqrt{\sigma^2(P)\sigma^2(Q)}} \quad (12)$$

At which, σ denotes variance. The existing expression for PCC ρ is signified by:

$$\rho = \frac{\mathrm{E(PQ)} - \mathrm{E(P)E(Q)}}{\sqrt{\sigma^2(P)\sigma^2(Q)}} \quad (13)$$

$$\rho = \frac{\sum(p_i - \overline{p_i})(q_i - \overline{q_i})}{\sqrt{\sum(p_i - \overline{p_i})^2 \sum(q_i - \overline{q_i})^2}} \quad (14)$$

$\overline{p_i}\ denotes\ average\ of\ P$, and mean of Q is $\overline{q_i}$.

The manipulation of local coefficient is carried using weighted PCC (W-PCC) in the specific position on including weights w_{xy} as given in below expression:

$$\mathrm{WPCC} = \frac{\sum w_{xy}\ (p_i - \overline{p_i})(q_i - \overline{q_i})}{\sqrt{\sum w_{xy}\ (p_i - \overline{p_i})^2 \sum (q_i - \overline{q_i})^2}} \quad (15)$$

The weights w_{ij} expression will be estimated by means of bi square function as:

$$w_{xy} = \begin{cases} \left[1 - \left(\frac{d_{xy}}{h_x}\right)^2\right]^2 & \text{if } d_{xy} < h_x \\ 0 \quad otherwise \end{cases} \quad (16)$$

In this, d_{xy} denotes the distance among location x and y. The selected bandwidth is given by h_x. From this, the rank is set for citation and the most relevant or suitable citation is recommended to the user.

4. PERFORMANCE ANALYSIS

An assessment is carried for proposed model and the outcomes attained are compared with existing models to validate the proposed model efficiency. For the performance assessment, the publicly available dataset is considered.

4.1 Details of Dataset considered

From the considered publicly accessible dataset, following are the details. It covers publication list of 50 researchers (Sugiyama & Kan, 2015) whose research interest are from several areas of computer science that ranges from security, software engineering, information retrieval, user interface, operating system, embedded system, database, and programming languages. The entire form of reference and citation will be retrieved from google scholar and all other papers that cites any one of references additionally to references of entire target's paper citation. Some statistics of employed dataset is given below in table 1:

Table 1. Statistics of input data considered.

Total number of researchers	50
Total number of papers recommended	100,351
Average number of researcher's publications	10
Average number of citations in paper recommended	17.9 (maxim-175)
Average number of citations of all researcher's publication	14.8 (maxim-169)
Average number of references in recommended papers	15.5 (maxim-53)
Average number of references of all researcher's publication	15.0 (maxim-58)

4.2 Estimation of Performance and Comparative analysis

The suggested scheme performance is estimated for various metrics such as Precision, recall, MRR, F-Measure, and MAP. The attained outcomes are related with various state-of-the-art schemes (Velkumar, 2022) such as collaborative, CCF, and the extraction model based on synthetization.

The evaluation of precision is carried with the use of subsequent expression:

$$Prec = \frac{\sum(relevant\ papers) \cap \sum(retrieved\ papers)}{\sum(retrieved\ papers)} \quad (17)$$

A metric Precision is measured for the suggested scheme and outcomes gained are related with several traditional techniques such as collaborative, CCF, and the extraction model based on synthetization to validate the performance efficiency of proposed model over others. From the plotted values, it is apparent that the precision rate of proposed model is higher than other models as revealed from figure 4 and table 2.

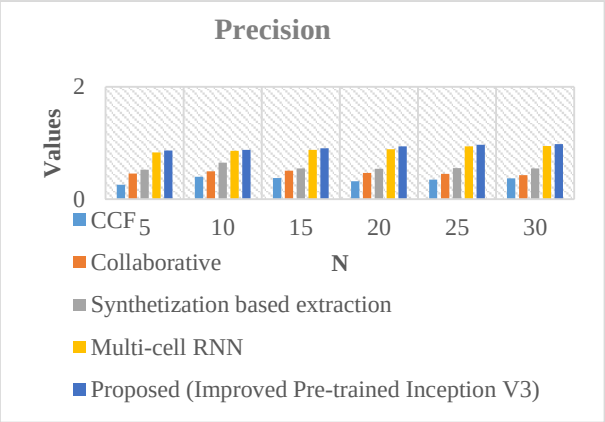

Figure 4. Precision comparative estimate

Table 2. Precision comparative estimate.

N	CCF	Collaborative	Synthetization based extraction	Multi-cell RNN	Proposed (Improved Pre-trained Inception V3)
5	0.26	0.46	0.525	0.834	0.87
10	0.4	0.5	0.65	0.862	0.88
15	0.375	0.51	0.55	0.877	0.91
20	0.32	0.47	0.54	0.889	0.94
25	0.35	0.45	0.552	0.94	0.97
30	0.37	0.43	0.551	0.945	0.98

The evaluation of recall is carried with the use of subsequent expression:

$$Rec = \frac{\sum(relvnt\ papers) \cap \sum(retrvd\ papers)}{\sum(retrvd\ papers)} \quad (18)$$

An analysis of recall is done for the suggested scheme and outcomes gained are related with several traditional techniques such as collaborative, CCF, and the extraction model based on synthetization to validate the performance efficiency of proposed model over others. From the plotted values, it is apparent that the recall rate of proposed model is higher than other models as revealed from figure 5 and table 3.

The evaluation of F1-measure is carried with the use of subsequent expression:

$$F1 - measure = \frac{\sum(relvnt\ papers) \cap \sum(retrvd\ papers)}{\sum(retrvd\ papers)} \quad (19)$$

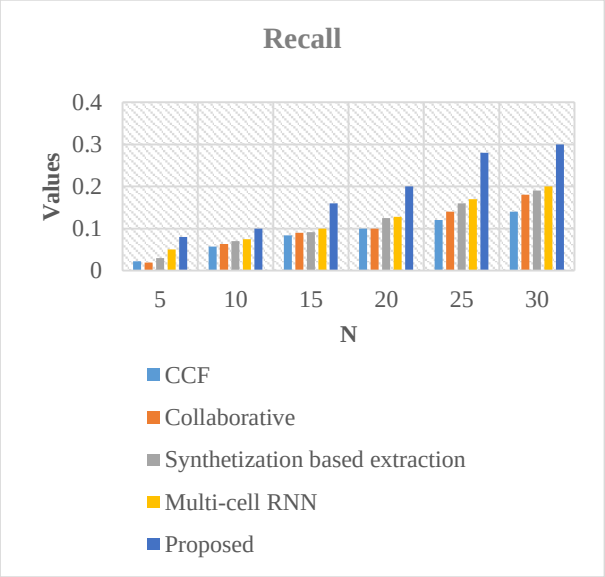

Figure 5. Recall comparative estimate

Table 3. Recall comparative estimate.

N	CCF	Collaborative	Synthetization based extraction	Multi-cell RNN	Proposed (Improved Pre-trained Inception V3)
5	0.022	0.019	0.03	0.05	0.08
10	0.057	0.063	0.07	0.075	0.1
15	0.084	0.09	0.091	0.1	0.16
20	0.1	0.1	0.125	0.128	0.2
25	0.12	0.14	0.16	0.17	0.28
30	0.14	0.18	0.19	0.2	0.3

An analysis of F-measure is done for the suggested scheme and outcomes gained are related with several traditional techniques such as collaborative, CCF, and the extraction model based on synthetization to validate the performance efficiency of proposed model over others. From the plotted values, it is apparent that the F-measure rate of proposed model is higher than other models as revealed from figure 6 and table 4.

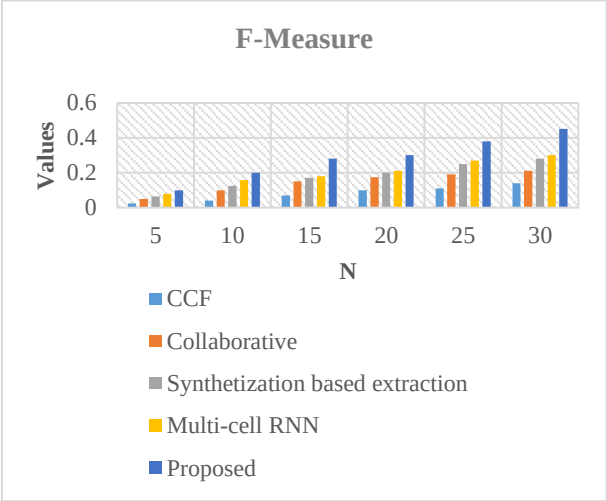

Figure 6. F-Measure comparative estimate

Mean Reciprocal Rank (MRR) is employed to measure the ability of system on effective retrieval of research papers that are at top rank in recommendations list. The expression representing this estimation are provided by:

$$MRR = \frac{1}{N_p}\sum_{i\in I}\frac{1}{rank(i)} \quad (20)$$

At which, rank (i) signifies rank at which i is identified that is relevant initial paper is recognized and target papers number are signified as N_p.

Table 4. F-Measure comparative estimate.

N	CCF	Collaborative	Synthetization based extraction	Multi-cell RNN	Proposed (Improved Pre-trained Inception V3)
5	0.025	0.05	0.065	0.079	0.1
10	0.04	0.1	0.125	0.158	0.2
15	0.07	0.15	0.17	0.18	0.28
20	0.099	0.175	0.2	0.21	0.3
25	0.11	0.19	0.25	0.27	0.38
30	0.14	0.21	0.28	0.3	0.45

The metric MRR is measured for the suggested scheme and outcomes gained are related with several traditional techniques such as collaborative, CCF, and the extraction model based on synthetization to validate the performance efficiency of proposed model over others. From the plotted values, it is apparent that the MRR value is higher than other compared models as of figure 7 and table 5.

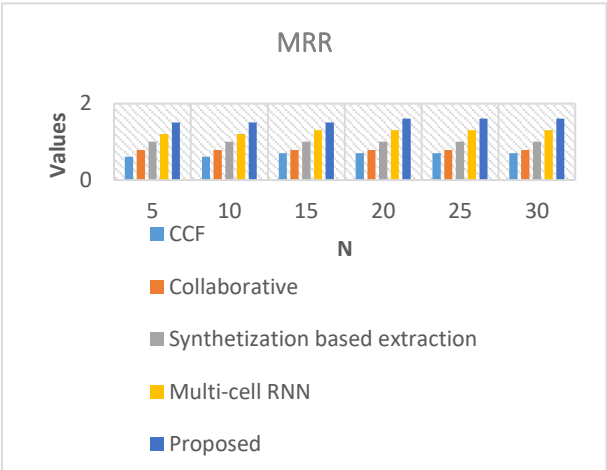

Figure 7. MRR comparative estimate

Mean Average Precision (MAP) metric is employed to measure the ability of system on effective retrieval of research papers that are at top rank in recommendations list. The expression representing this estimation are provided as shown:

$$MAP = \frac{1}{I}\sum_{i\in I}\frac{1}{ni}\sum_{k=1}^{N}P(R_{ik}) \quad (21)$$

At which, rank (i) signifies rank at which i is identified that is relevant initial paper is recognized and target papers number are signified as N_p.

Table 5. MRR comparative estimate.

N	CCF	Collaborative	Synthetization based extraction	Multi-cell RNN	Proposed (Improved Pre-trained Inception V3)
5	0.61	0.78	1.0	1.2	1.5
10	0.61	0.78	1.0	1.2	1.5
15	0.7	0.78	1.0	1.3	1.5
20	0.7	0.78	1.0	1.3	1.6
25	0.7	0.78	1.0	1.3	1.6
30	0.7	0.78	1.0	1.3	1.6

An analysis of MRR is made for the suggested scheme and outcomes gained are related with several traditional techniques such as collaborative, CCF, and the extraction model based on synthetization to validate the performance efficiency of proposed model over others. From the plotted values, it is apparent that the MAP value is higher than other compared models as of figure 8 and table 6.

On analyzing the attained outcomes, the proposed model proves their superiority over other existing models that are compared related to the citation recommendation schemes. Henceforth, the relevant and suitable papers are recommended for researchers to make use of effective citation records for their research articles, documentations, and so on.

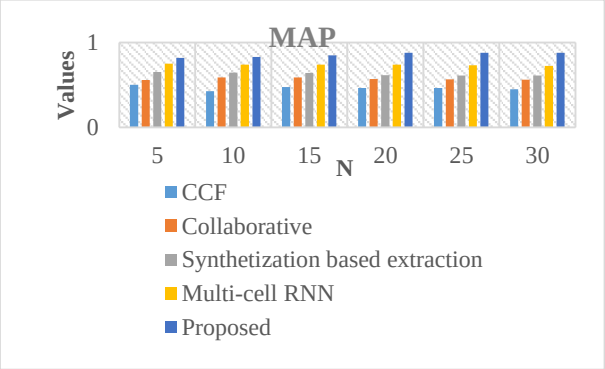

Figure 8. MAP comparative estimate

Table 6. MAP comparative estimate.

N	CCF	Collaborative	Synthetization based extraction	Multi-cell RNN	Proposed (Improved Pre-trained Inception V3)
5	0.501	0.558	0.651	0.752	0.82
10	0.425	0.59	0.645	0.741	0.83
15	0.475	0.588	0.64	0.741	0.85
20	0.465	0.57	0.615	0.738	0.88
25	0.465	0.565	0.613	0.733	0.88
30	0.449	0.561	0.61	0.726	0.88

5. CONCLUSION

A framework based on transfer learning model is employed in this article to make effective citation recommendation system. Primarily, input dataset was considered and stop words or functions words removal were performed using stochastic pattern matching algorithm through which the library creation was carried. Then, document canonicalization was performed by means of Term Frequency-Inverse Document Frequency (TF-IDF) after which the stored data are predicted and classified by means of Improved Pre-trained Inception V3 classifier. Before classification process, hyper-parameters were tuned on employing MP-HPO model. This in turn enhances the classifier performance aiding in effective recommendation of citation records. Eventually, ranking was done by means of employing weighted Pearson correlation coefficient to rank the predicted citation and recommend best ones by means of ranking approach. The designed proposed technique is validated on various metrics like precision, MAP, MRR, Recall, and F1-score. The proposed outcomes are compared with various existing schemes to show the enhancement of proposed strategy over others. The validated outcome reveals the superiority of suggested scheme over other compared existing models. The findings made suggested that the incorporation of transfer learning along with optimization model aids in dealing with various issues faced by huge range of available data in the library of digital academics. Specifically, the optimization process with TL classifier upgrades the system performance highly. As a future work, this system could be improved on making it as distributed or parallel system, thus running in various number of GPU's or machines. Also, for canonicalization models, more effective models such as BERT, Auto-encoder (AE) and so on will be employed rather than TF-IDF will employed for generating low-dimensional embedding of text.

References:

Abbas, M. A., Ajayi, S., Bilal, M., Oyegoke, A., Pasha, M., & Ali, H. T. (2024). A deep learning approach for context-aware citation recommendation using rhetorical zone classification and similarity to overcome cold-start problem. *Journal of Ambient Intelligence and Humanized Computing, 15*(1), 419-433.

Bagul, D. V., & Barve, S. (2021). *A novel content-based recommendation approach based on LDA topic modeling for literature recommendation*. Paper presented at the 2021 6th International conference on inventive computation technologies (ICICT).

Chen, W., Shen, Z., Pan, Y., Tan, K., & Wang, C. (2024). Applying machine learning algorithm to optimize personalized education recommendation system. *Journal of Theory and Practice of Engineering Science, 4*(01), 101-108.

Devika, P., & Milton, A. (2025). Book recommendation using sentiment analysis and ensembling hybrid deep learning models. *Knowledge and Information Systems, 67*(2), 1131-1168.

Gündoğan, E., & Kaya, M. (2022). A novel hybrid paper recommendation system using deep learning. *Scientometrics, 127*(7), 3837-3855.

Gündoğan, E., Kaya, M., & Daud, A. (2023). Deep learning for journal recommendation system of research papers. *Scientometrics, 128*(1), 461-481.

Huang, W., Liu, B., & Wang, Z. (2023). Paper Recommendation via Correlation Pattern Mining and Attention Mechanism. *Journal of Sensors, 2023*(1), 3311363.

Huang, Y., Wang, H., & Wang, R. (2022). Deep learning recommendation algorithm based on semantic mining. *Plos one, 17*(9), e0274940.

Huynh, S. T., Dang, N., Nguyen, D. H., Huynh, P. T., & Nguyen, B. T. (2023). FPSRS: a fusion approach for paper submission recommendation system. *Applied Intelligence, 53*(8), 8614-8630.

Jeong, C., Jang, S., Park, E., & Choi, S. (2020). A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics, 124*, 1907-1922.

Jie, C., Yang, L., Shu, Z., & Yanping, Z. (2021). Citation recommendation via hierarchical attributed network representation learning. *Journal of Frontiers of Computer Science & Technology, 15*(6), 1103.

Ko, H., Lee, S., Park, Y., & Choi, A. (2022). A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics, 11*(1), 141.

Li, W. (2024). Scientific paper recommender system using deep learning and link prediction in citation network. *Heliyon, 10*(14).

Li, Y., Liu, K., Satapathy, R., Wang, S., & Cambria, E. (2024). Recent developments in recommender systems: A survey. *IEEE Computational Intelligence Magazine, 19*(2), 78-95.

Liang, Y., & Lee, L.-K. (2023). A Systematic Review of Citation Recommendation Over the Past Two Decades. *International Journal on Semantic Web and Information Systems (IJSWIS), 19*(1), 1-22.

Lin, J., Dai, X., Xi, Y., Liu, W., Chen, B., Zhang, H., . . . Zhu, C. (2025). How can recommender systems benefit from large language models: A survey. *ACM Transactions on Information Systems, 43*(2), 1-47.

Ma, A., Liu, Y., Xu, X., & Dong, T. (2021). A deep-learning based citation count prediction model with paper metadata semantic features. *Scientometrics, 126*(8), 6803-6823.

Ma, S., Zhang, C., Zhang, H., & Gao, Z. (2025). Citation recommendation based on argumentative zoning of user queries. *Journal of Informetrics, 19*(1), 101607.

Roy, D., & Dutta, M. (2022). A systematic review and research perspective on recommender systems. *Journal of Big Data, 9*(1), 59.

Sugiyama, K., & Kan, M.-Y. (2015). A comprehensive evaluation of scholarly paper recommendation using potential citation papers. *International Journal on Digital Libraries, 16*(2), 91-109.

Urdaneta-Ponte, M. C., Mendez-Zorrilla, A., & Oleagordia-Ruiz, I. (2021). Recommendation systems for education: Systematic review. *Electronics, 10*(14), 1611.

Velkumar, B. (2022). Deep Learning-Assisted Citation Recommendation System Using Multi- cell RNN Approach. *Tuijin Jishu/Journal of Propulsion Technology, 44*. doi:10.52783/tjjpt.v44.i2.135.

Wang, J., Zhu, L., Dai, T., & Wang, Y. (2020). Deep memory network with bi-lstm for personalized context-aware citation recommendation. *Neurocomputing, 410*, 103-113.

Wei, B. (2024). Paper Recommendation Method based on Attention Mechanism and Graph Neural Network. *J. Electrical Systems, 20*(2), 88-95.

Xie, Q., Zhu, Y., Huang, J., Du, P., & Nie, J.-Y. (2021). Graph neural collaborative topic model for citation recommendation. *ACM Transactions on Information Systems (TOIS), 40*(3), 1-30.

Yang, D., Wang, H., Shi, Z., & Zhu, K. (2024). A deep learning approach to enhance accuracy and diversity of recommendation for interdisciplinary journals.

Yuan, M., Wan, J., & Wang, D. (2023). *CRM-SBKG: Effective citation recommendation by siamese BERT and knowledge graph*. Paper presented at the 2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA).

Zhang, J., & Zhu, L. (2022). Citation recommendation using semantic representation of cited papers' relations and content. *Expert Systems with Applications, 187*, 115826.

Zhang, Q., Lu, J., & Jin, Y. (2021). Artificial intelligence in recommender systems. *Complex & Intelligent Systems, 7*(1), 439-457.

Velkumar K
Department of Computer Science and Engineering, Kalasalingam Academy of Research Education, Krishnankovil, Tamilnadu, India
Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Green Fileds, Vaddeswaram, A.P. – 522302, India
velkumarskc@gmail.com
ORCID 0009-0006-7972-6489

Balasubramanian Chokkalingam
Department of Computer Science and Engineering, Kalasalingam Academy of Research Education, Krishnankovil, Tamilnadu, India
baluece@gmail.com
ORCID 0000-0002-3721-7743