Afr. J. Lib, Arch. and Inf. Sc. Vol. 35, No. 2 (July 2025) 275-288
https://doi.org/10.5281/zenodo.20751708

Digital Literacy and Knowledge Management in Libraries: AI-Driven Insights into Information Retrieval, User Behavior, and Mental Health

Chengzhe Guo
Faculty of Education, Beijing Normal University, 100875, China
guochengzhe@bnu.edu.cn

and

Shangjie Meng*
School of Automation, Chongqing University of Posts and Telecommunications, 400065, China
shangjie_meng@163.com

Abstract
Libraries have transformed into advanced learning spaces where digital literacy and knowledge management are central to enhancing user engagement and overall well-being. Conventional library analytics, however, are limited in their capacity to interpret complex patterns of user behaviour and emotional cues that affect both information-seeking processes and mental health outcomes. This study aims to develop an artificial intelligence (AI)-based predictive framework for academic libraries, designed to strengthen digital literacy initiatives, optimise knowledge management, and provide detailed insights into user behaviour and psychological states. User interaction data were sourced from digital library platforms, including search query logs, access frequencies, and feedback submissions. Incomplete and redundant entries were removed, and textual data were converted into numerical representations using the Term Frequency-Inverse Document Frequency (TF-IDF) method to facilitate computational analysis. The Pearson Correlation Coefficient (PCC) was employed to examine associations among information retrieval patterns, cognitive load features, and emotional responses. The Bitterling Fish Optimizer tuned Decision Forest Tree (BFO-DFT) model, a hybrid machine learning (ML) approach, integrates Bitterling Fish Optimization (BFO) for feature selection with Decision Tree (DT) and Random Forest (RF) algorithms for classification, enabling predictions of user satisfaction and potential stress levels. Findings indicate that the BFO-DFT model demonstrates robust predictive performance and generalisation capacity. Specifically, the model achieved superior metrics with an accuracy of 0.83, recall of 0.87, and F1-score of 0.85 in forecasting user behaviour and satisfaction. Overall, the results highlight that AI-driven optimisation can enhance information retrieval efficiency, foster digital literacy development, and facilitate the early identification of mental stress indicators, thereby supporting intelligent and sustainable library environments.

Keywords: Knowledge Management, User Engagement, Library, Mental Health, Cognitive Load, Emotional Attitude.

Introduction

Digital literacy and knowledge management (KM) are pivotal in contemporary library environments (Rafi et al., 2022). Digital literacy refers to the capability of individuals to locate, assess, and utilise digital information across various technical platforms, emphasising information ethics, critical thinking, and the effective application of digital tools (Samia and Hind, 2024). KM within libraries involves the systematic organisation, dissemination, and utilisation of information to support decision-making and learning processes (Weck et al., 2021). Both digital literacy and KM share objectives such as improving information accessibility, promoting lifelong learning, and facilitating users' cognitive and emotional development (Poquet and de Laat, 2021). Effective implementation of these practices fosters a participatory learning environment that interacts seamlessly with information systems and encourages active engagement with digital content (Camacho et al., 2021).

Prior to the adoption of AI and advanced data analytics, libraries relied on manual and rule-based approaches to manage and interpret user interactions (Nasir, 2024). Traditional methods included catalogue-based retrieval systems and basic keyword search engines with minimal contextual understanding (Zhang et al., 2023). Boolean logic and metadata indexing were central to information retrieval, offering structured but rigid access to library resources. Knowledge was maintained through manual documentation, librarian-led curation, and elementary database management systems. User activity tracking was limited to surveys, circulation records, and access logs, leaving behavioural and emotional dimensions—such as mental health or affective engagement—unexplored (Balcombe and De Leo, 2022). The AI-powered framework for an Intelligent Library Ecosystem is illustrated in Figure 1.

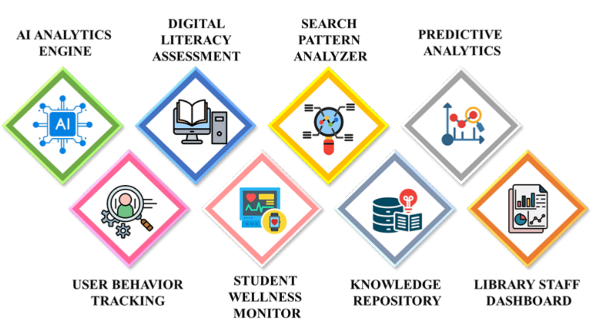

Figure 1: AI-Driven Library Intelligence Framework for Digital Literacy and Knowledge Management.

Traditional methods provided a structured approach for accessing and storing information; however, they were constrained by their inability to process complex, unstructured data, limited adaptability, and minimal customization options (Malekloo et al., 2022). These limitations prompted the shift towards AI-driven intelligent systems to manage and enhance the overall user experience, revealing an informational gap regarding the influence of user behaviour, digital engagement, and emotional states on learning and well-being in library contexts (Aldoseri et al., 2024). AI-based information retrieval systems offer the potential to address scalability and adaptability challenges when handling large, unstructured datasets across multiple library platforms. The objective is to enhance accuracy, personalisation, and accessibility by designing an intelligent information retrieval and recommendation system grounded in AI and the Bacterial Foraging Optimization-Discrete Fourier Transform (BMO-DFT) framework within digital library systems. The principal contributions of this research are as follows:

- Dataset Usage: Real-world user interactions, search behaviours, and emotional indicators were incorporated using the Library Behaviour and Mental Health Dataset from Kaggle.
- Preprocessing and Feature Extraction: Data were cleaned and represented using TF-IDF and semantic clustering, with key textual features extracted to improve retrieval performance.
- Optimisation Process: The BFO algorithm was employed to optimise feature weights, thereby enhancing both retrieval accuracy and system flexibility. A scalable AI system was developed, applicable to academic and public libraries for real-time information retrieval and monitoring of mental health.
- Simulation Results: Improvements in simulation conditions, reflected in F1-score, precision, and recall, contributed to more accessible digital library systems.

The structure of the research paper is as follows: Section 1 introduces the study, outlining the research aims and objectives. Section 2 presents a review of related work on AI-based library systems. The proposed methodology is detailed in Section 3. Section 4 discusses the results and analysis of experimental findings. Finally, Section 5 provides the conclusion and directions for future research.

Related Works

Insufficient knowledge of AI and digital literacy hindered school principals from effectively managing institutions and implementing simulations, seminars, and real-world technological applications centred on AI-based digital literacy (Rohayani et al., 2024). Enhancing comprehension and integration of AI technologies into administrative and educational processes is essential for improving the capacity to address challenges in the era of digital learning. Barman (2024) examined applications of ML and AI, including automation, predictive analytics, and personalised recommendations. Libraries face challenges such as managing large volumes of digital data, boosting user engagement, and enhancing operational efficiency. While AI and ML facilitate accessibility and efficiency in libraries, issues such as algorithmic bias, data privacy, and workforce displacement remain significant. User profiling techniques have been applied to analyse multiple online actions, including information seeking, verification, and sharing (Kanwal et al., 2024a). ML methods, particularly K-Means clustering, were employed on multidimensional behavioural data to identify user intent profiles. The model effectively categorised diverse user intent types, validated through human evaluation.

Despite the expansion of online healthcare, only a few initiatives successfully promote patient digital literacy and engagement (Alon et al., 2023). The Digital Outreach for Obtaining Resources and Skills (DOORs) programme, developed over five years through community partnerships, curriculum updates, LMS-based online learning, and standardised training, was assessed using both evaluation metrics and semi-structured approaches. Participants reported increased confidence in 72% of assessed online skills. The DOORs model has contributed to skill enhancement, patient-centric design, and improved understanding of the digital divide. In Technology Enhanced Learning (TEL), semantically relevant Learning Objects (LOs) from diverse repositories are accessed (Tahir et al., 2022). The Dynamic Recommendation of Filtered Learning Objects (DRFLO) system integrates ML with context-aware filtering to improve LO recommendation and ranking. The DRFLO framework was applied to retrieve intelligent LOs and optimise course planning, achieving 93% accuracy.

A user intent ML system, combining K-Means clustering for profiling and Linear Discriminant Analysis Classifier (LDAC) for classification, was developed by Kanwal et al. (2024b). As unfiltered hybrid data grows, limited information is available regarding users' online activities such as searching, sharing, and verifying information. The system attained 80% predictive accuracy and 67% validation accuracy in classifying users based on their online information behaviours. Digital libraries have been enhanced to provide inclusive and personalised knowledge access to a broad range of users (Ikwuanusi et al., 2023). AI technologies, including chatbots, recommender systems, ML, and NLP, were deployed to assess accessibility, user engagement, and information delivery. Challenges such as algorithmic bias and data privacy persist, yet AI-assisted solutions have improved personalisation and service accessibility.

E-learning participants contribute to the identification of relevant and current learning materials, mitigating issues caused by unreliable search results and information overload (Shahbazi and Byun, 2022). A virtual intelligent agent-based recommendation system, utilising ML clustering, NLP, and semantic analysis, was developed to examine user profiles and preferences. The system enhanced predictive accuracy to 98%, while improving content relevance and customisation in e-learning platforms. Social networks and recommendation systems are inherently complex and pose challenges in simulating realistic user behaviours (Wang et al., 2025). LLM-based agents, equipped with profiling, memory, and action modules, were developed to emulate human behaviours and interactions. Personalised control and social diversity in agent design helped mitigate issues related to incompetence and conformity.

Accurately determining the academic performance of online learners relies on analysis of their digital activity data (Qiu et al., 2022). A self-adaptive feature fusion algorithm, integrating entropy weighting and variance filtering with the E-learning Behaviour Classification (EBC) Model, was introduced to identify effective learning behaviour attributes and enhance predictive precision. The approach significantly supported instructors in forecasting learning outcomes. The Deep Personalized Course Recommendation System (DORIS) was implemented to align course offerings with student preferences. Traditional systems faced challenges due to information overload and scaling issues, leading to incorrect course selection; DORIS demonstrated higher relevance and practical alignment compared with existing recommendation methods (Ma et al., 2023).

Massive Open Online Courses (MOOCs) present challenges in guiding learners' course and learning path selections, potentially affecting performance (Amin et al., 2023). A smart e-learning platform, leveraging RL, Markov Decision Processes (MDP), and Q-learning, was proposed to provide adaptive and flexible learning path suggestions (Ezaldeen et al., 2022). A DL system integrating Deep Belief Networks (DBN) for text summarisation and image captioning with BiLSTM for information retrieval was developed. Traditional algorithms relying on large labelled datasets and manual feature engineering limit effectiveness and flexibility (Romero-Ochoa et al., 2025). The Bilingual Evaluation Understudy (BLEU) metric showed high precision and recall (Ayman and Abo El Rejal, 2023). Neutrosophic fuzzy sets were employed within a Neutrosophic Convolutional Neural Network (NCNN) to manage multimodal inputs (text and image) for feature fusion and uncertainty handling. Conventional single-modality methods perform poorly under noisy conditions (Wajid et al., 2024). The NCNN outperformed single-modality approaches in terms of accuracy and stability on multimodal classification datasets.

On social media platforms, the spread of false or misleading information disrupts users and hinders effective information retrieval (Thaher et al., 2021). The BHHO-LR model, integrating NLP, feature selection, BoW, and TF-IDF, achieved a 5% improvement in information retrieval accuracy compared with prior methods. Collaborative Filtering (CF) combined with Content-Based Filtering (CBF) enhanced semantic clustering and identification of meaningful user-book relationships. Traditional recommendation algorithms were prone to inaccuracy due to data bias and sparse item-user interactions (Wayesa et al., 2023). Evaluations of precision, recall, and F-measure demonstrated that contemporary models were more effective in providing recommendations. Educational resource management in libraries presents challenges for efficient retrieval (Ma, 2024). ML algorithms incorporated in the CATALYST platform improved forecasting and retrieval rates, enhancing resource accessibility, inventory management, and students' comprehension of course content.

Problem Statement

The limitations inherent in AI-driven library frameworks impact fairness and ethical implementation, including issues such as algorithmic bias, data privacy concerns, and insufficient transparency within ML and NLP techniques (Shahbazi and Byun, 2022). Similarly, the scalability and adaptability of dynamic e-learning environments can be constrained by weaknesses in self-adaptive feature fusion strategies, such as overfitting, limited generalisation across heterogeneous datasets, and dependence on predefined behavioural categories (Qiu et al., 2022). The proposed BFO-DFT approach mitigates these limitations by optimising feature selection, enhancing transparency, reducing bias, and providing improved scalability and flexibility when handling diverse datasets within AI-driven library and e-learning ecosystems.

Methodology

The proposed approach examines user interaction patterns and psychological well-being within digital libraries using the Library Behaviour and Mental Health Dataset (Kaggle). Textual data are preprocessed and transformed into TF-IDF representations, while the PCC is employed to identify pertinent features. The BFO-DFT hybrid model, combining Bitterling Fish Optimizer and Decision Forest Tree, facilitates enhanced feature selection and classification. This integration of information renders library management systems more intelligent and adaptive, improving information retrieval, forecasting user satisfaction, and detecting indicators of mental stress. The operational workflow of the BFO-DFT algorithm is illustrated in Figure 2.

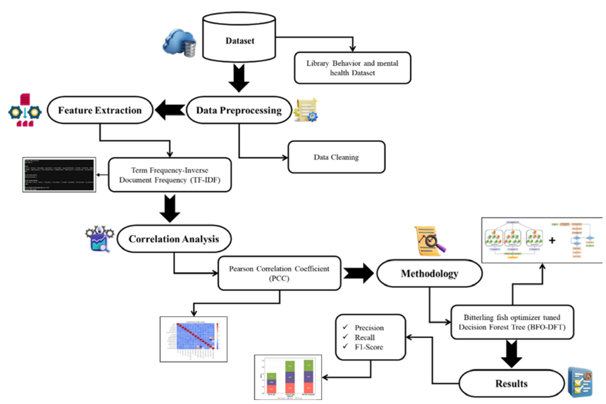

Figure 2: An Overview of the BFO-DFT Algorithm's Flowchart.

Data Cleaning for Preprocessing

The prediction model, grounded in AI and implemented within library systems, employed rigorous data cleaning to ensure high-quality inputs.

Redundant, duplicate, and incomplete records of user interactions, including search logs, access frequencies, and feedback entries, were removed. Inconsistent textual formats were standardised, and missing values were imputed using the mode. Outliers affecting behavioural and emotional metrics were eliminated based on statistical thresholds. The resulting dataset, reflecting regular user activity patterns, enabled accurate estimation of digital literacy, evaluation of information retrieval efficiency, and early identification of indicators of mental stress. An overview of data cleaning and encoding for the Library Behaviour Dataset is presented in Figure 3.

```
✅ Dataset Loaded Successfully!
Shape: (1790, 15)

Columns:
 ['user_id', 'session_id', 'search_queries', 'query_diversity', 'access_frequency', 'avg_session_duration_min', 'click_depth', 'doc_downloads', 'feedback
_sentiment', 'cognitive_load_score', 'info_retrieval_efficiency', 'collaboration_index', 'emotion_tone_score', 'stress_level_label', 'user_satisfaction_l
abel']

✅ Data Cleaning Completed!
Remaining Missing Values: 0

✅ Label Encoding Completed!
Encoded Columns: ['user_id', 'session_id', 'search_queries', 'access_frequency', 'click_depth', 'doc_downloads', 'stress_level_label', 'user_satisfaction
_label']

✅ Top Strongly Correlated Feature Pairs (|r| > 0.75):
Series([], dtype: float64)
```

Figure 3: Dataset Preprocessing Workflow Summary.

Term Frequency–Inverse Document Frequency (TF-IDF)

TF-IDF is employed to extract key textual features from user interaction data, supporting the enhancement of the AI-driven prediction framework for accurate information retrieval and mental health assessment within digital library systems. This method effectively transforms user query submissions into corresponding feature vectors. TF-IDF calculates the weight of each term, representing its overall significance across the dataset. The importance of user interaction terms in relation to digital literacy and mental health is formalised in Equation (1).

$$TF - IDF(s,c) = TF(s,c) \times IDE(s) \quad (1)$$

The inverse document frequency (IDF) is defined in Equation (2), where IDF represents the inverse document frequency, and TF(s,c) denotes the term frequency of token (s) within document (c).

$$IDF(s) = log\,\frac{1+m}{1+DF(s)} + 1 \quad (2)$$

The inverse document frequency is denoted as IDF(s), where m represents the total number of documents in the collection, and DF(s) indicates the number of tweets containing the term s. The TF-IDF vector (U) for each tweet document is subsequently normalised using the L2-norm, as defined in Equation (3).

$$U = \frac{U}{\sqrt{U_1^2 + U_2^2 + U_m^2}} \quad (3)$$

Here, U represents the normalised feature vector, and U_1, U_2, …, U_m correspond to the individual feature components. This approach efficiently transforms textual user data into meaningful numerical features, enabling accurate prediction of behavioural patterns and mental health indicators.

3.2 Pearson Correlation Coefficient (PCC)

Within digital library interaction data, the PCC is employed to identify relationships between user behaviours, cognitive load, and emotional indicators. The PCC, originally developed by Pearson, quantifies both the strength and direction of the association between two variables. Equation (4) represents the calculation of PCC between variables z (mental or retrieval outcome) and w_j (user behavior features):

$$q_{w,z} = \frac{\sum_{j=1}^{m}(w_j - w)(z_j - z)}{\sum_{j=1}^{m}(w_j - w)^2 \sum_{j=1}^{m}(z_j - z)^2} \quad (4)$$

Where w_j is the -th value for the observational variable w, z_j the -th value of variable z, the correlation strength is represented by the absolute value of $q_{w,z}$, which ranges from +1 to -1. In this case, $q_{w,z} = 1$ shows a perfect positive linear correlation, $q_{w,z} = 1$ indicates a perfect negative linear correlation, and $q_{w,z} = 0$ signifies no linear association. The PCC was applied to assess the relationships among emotional responses, cognitive load indicators, and information retrieval behaviours. For subsequent ML analysis, features exhibiting an absolute PCC value greater than 0.5 were selected as significant inputs. An overview of the correlation analysis between user behaviour and satisfaction metrics is presented in Figure 4.

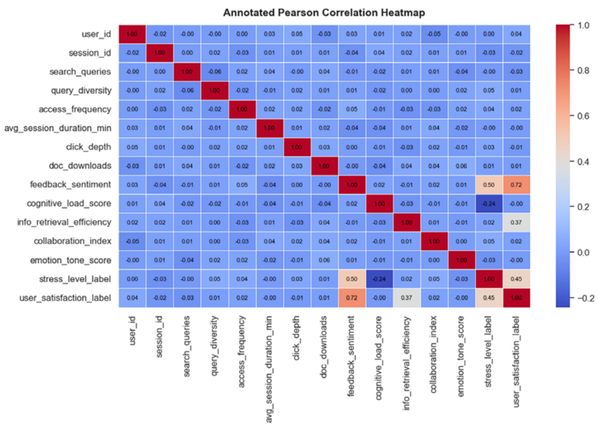

Figure 4: Pearson Correlation Heatmap of Library User Behavior and Mental Health Indicators.

Bitterling Fish Optimizer Tweaked Decision Forest Tree (BFO-DFT)

The BFO-DFT model enhances the prediction of user behaviour and mental health within digital library systems by integrating evolutionary intelligence with ensemble learning. BFO simulates the competitive reproductive behaviour of bitterling fish, balancing exploration and exploitation to achieve optimal feature selection. The Decision Forest Tree, which combines the interpretability of DT with the predictive accuracy of RF, is employed to classify the selected features. Through adaptive learning and intelligent, data-driven optimisation, the proposed BFO-DFT approach improves early detection of stress indicators, prediction of digital literacy, and efficiency of information retrieval. The workflow of the BFO-DFT optimisation algorithm for intelligent library analysis is detailed in Algorithm 1.

The hyperparameters employed to optimise the proposed BFO-DFT model are presented in Table 1. The configuration includes ten selected features, a population size of thirty, and a maximum of one hundred iterations. The exploration and exploitation coefficients are set within ranges of 0.2–0.8 and 0.1–0.5, respectively. Feature selection is guided by a PCC threshold of 0.5, and the fitness function is defined as (1 / (Acc + F1 + Rec)) to achieve a balance between model robustness and predictive accuracy.

Algorithm 1: BFO-DFT

Input:
D ← Library Behavior and Mental Health Dataset (from Kaggle)
M ← 30
C ← 10
S_max ← 100
α_range ← [0.2, 0.8]
β_range ← [0.1, 0.5]
threshold_PCC ← 0.5
Output:
W_best ← Optimal feature weight vector
Model ← Trained Decision Forest Tree
Step 1: Data Preprocessing
1.1 *Load dataset D*
1.2 *Remove duplicates, missing values → Impute mean/mode*
1.3 *Normalize text data using TF − IDF:*
$U = U / \sqrt{(U_1^2 + U_2^2 + \ldots + U_m^2)}$
1.4 *Compute Pearson Correlation Coefficient (PCC):*
$q(w,z) = \Sigma[(w\_j - \bar{w})(z\_j - \bar{z})] / \sqrt{(\Sigma(w\_j - \bar{w})^2 \Sigma(z\_j - \bar{z})^2)}$
1.5 *Select features where |q(w, z)| ≥ threshold_PCC*
Step 2: Initialize Bitterling Fish Population
2.1 *For j = 1 to M:*
Randomly initialize W_j = [w_j1, w_j2, …, w_jC]
2.2 *Initialize iteration counter s = 0*
2.3 *Define objective function e(W_j):*
e(W_j) = 1 / (Accuracy + F1_score + Recall)
Step 3: Evaluate Initial Fitness
3.1 *For each fish j:*
Evaluate fitness e(W_j)
3.2 *Set W_best = argmin(e(W_j))*
Step 4: Update Positions (Movement Rule)
While s < S_max:
For each fish j in the population:
α_j = random(α_range)
β_j = random(β_range)
W_rand = Randomly selected fish position
Update position using:
W_j^(s + 1) = W_j^s + α_j ∗ (W_best − W_j^s) + β_j ∗ (W_j^s − W_rand)
Evaluate new fitness e(W_j^(s + 1))
If e(W_j^(s + 1)) ≤ e(W_j^s):
W_j = W_j^(s + 1)
Else:
Retain old position
Update W_best = argmin(e(W_j))
s = s + 1
Step 5: Termination
If s ≥ S_max or no improvement in W_best for 10 iterations:
Terminate optimization
Return W_best
Step 6: Train Decision Forest Tree (DFT)
6.1 *Input selected features (based on W_best) to DFT*
6.2 *Train multiple Decision Trees:*
Each tree uses a subset of features & samples
6.3 *Combine results using Random Forest aggregation:*
RFfi_j = (Σ(normfi_ji)) / S
6.4 *Compute final feature importance and classification rules*
Step 7: Output and Evaluation
7.1 *Evaluate model performance:*
Accuracy, Precision, Recall, F1 − Score, Satisfaction rate
7.2 *Output:*
Optimal weights (W_best)
Trained a DFT model for prediction

Table 1: Hyperparameter Settings for BFO-DFT Model.

Parameter Name	Value / Range
Population Size	30
Number of Features	10
Maximum Iterations	100
Exploration Coefficient	[0.2, 0.8]
Exploitation Coefficient	[0.1, 0.5]
PCC Threshold	0.5
Fitness Function	1 / (Acc + F1 + Rec)
Decision Trees per Forest	100
Max Tree Depth	10
Minimum Samples per Split	2
Random State	42
Termination Condition	10 Iterations

Decision Tree (DT) for Predicting User Satisfaction

The DT algorithm enables structured prediction of information retrieval patterns and mental health indicators by initiating the classification of user behaviour. In contrast to earlier classification approaches, the DT model is interpretable and serves as an active learning mechanism. Oval-shaped leaf nodes indicate the final classification outcomes, while square-shaped internal nodes represent conditions or features that partition the data into multiple branches. Each leaf node, assigned a class label, is connected to the root node through a series of internal decision pathways. The path from the root to a leaf node establishes a classification rule. These rules were applied to analyse patterns in user behaviour, information retrieval efficiency, and emotional responses within the digital library system. Their interpretability allows librarians and system designers to monitor user engagement, information access, and mental health indicators in real time. The initial information entropy of dataset T is calculated using Equation (5).

$$Entropy\ (T) = -\sum_{b=1}^{n} o_b \times \log_2(o_b) \quad (5)$$

Where o_b is the percentage of a sample class and n is the total number of classes, T is the dataset. This process yields two potential outcomes:
Case 1: *Entropy* (T) = $log_2\ n$ *(highest)* if all the data are given unique class labels, $[o_b = \frac{1}{n}]$.
Case 2: *Entropy* (T) = *zero* (*lowest*) if all the data have the same class label, $[o_b = n = 1]$.

Divide T into (T_{left} *and* T_{right}), two attribute partitions.

The split entropy for each subset T is calculated using Equation (6).

$$Entropy_B(T) = \frac{|T_{left}|}{|T|} \times Entropy(T_{left}) + \frac{|T_{right}|}{|T|} \times Entropy(T_{right}) \quad (6)$$

Here, B represents an attribute of T. The number of samples in T, T_{left}, and T_{right}, respectively, is represented by $|T|$, $|T_{left}|$, and $|T_{right}|$.

The information gain associated with attribute B is calculated using Equation (7).

$$Information_{gain} = Entropy(T) - Entropy_B(T) \quad (7)$$

A higher $Information_{gain}$ number indicates greater entropy reduction, which improves a characteristic.

The C4.5 algorithm computes the split information value, as shown in Equation (8), to normalise information gain and mitigate the risk of overfitting.

$$Split_{info}(B) = -\sum_{b=1}^{l} \frac{|T_b|}{|T|} \times \log_2\left[\frac{|T_b|}{|T|}\right] \quad (8)$$

Where $|T|$ is the dataset's total instance, $|T_b|$ is the quantity of sub-datasets, *and* T_b, l is the number of subsets formed by attribute B.

The information gain ratio for each DT node is calculated using Equation (9).

$$IG_{ratio}(B) = \frac{Information_{gain}}{Split_{info}(B)} \quad (9)$$

Where the information gain ratio of attribute B is represented by IG_{ratio} (B). The characteristic with the highest IG_{ratio} value is chosen. This process is performed recursively to partition T into multiple refined subsets. By employing this approach, prediction of user behaviour and mental health trends within library systems becomes feasible, while ensuring that decision-making remains transparent and interpretable. The ensemble learning framework, utilising majority voting or averaging, is illustrated in Figure 5.

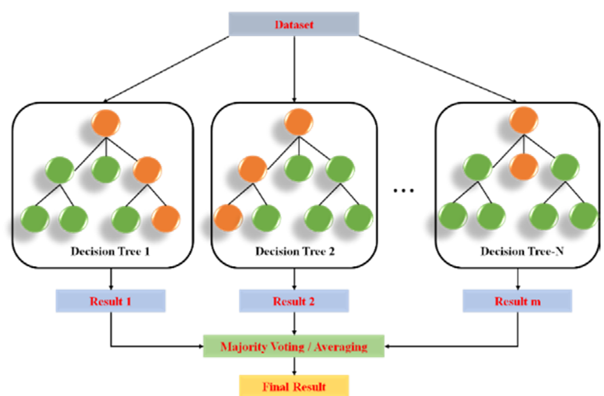

Figure 5: Architecture of BFO-DFT Ensemble Model for Stress Level Prediction.

Random Forest (RF) for Predicting Stress Levels

Within digital library systems, user activities were classified according to satisfaction and stress levels predicted by the RF algorithm. The RF method addresses both regression and classification tasks. For regression problems, as a forest comprises multiple DTs, the final output is calculated as the average of the predictions from all trees. For classification tasks, the predicted outcome is determined by the most frequent result among the constituent DTs, as formalised in Equation (10).

$$RFfi_j = \frac{\sum_{i \in alltrees} normfi_{ji}}{S} \quad (10)$$

Where $RFfi_j$ is the Random Forest feature importance score for feature j, $normfi_{ji}$ is the normalized feature importance of feature j within tree i; S is the total number of trees. The significance of each feature is assessed for every decision tree, and Equation (11) computes the overall feature importance as the average across all trees.

$$fi_j = \sum_{i:\, nodes\ i\ splits\ on\ feature\ j} t_i D_i \quad (11)$$

Where $\sum_{i:\ nodes\ i\ splits\ on\ feature\ j}$ represents the summation over all nodes i within the tree that utilise feature j for splitting. D_i is the decrease in impurity, *and* fi_j is the feature importance score for feature j. Samples that reach node i are represented by t sub i, and fi sub j is feature j: its importance. This ensures precise classification and reliable prediction of stress levels and user satisfaction within digital library systems.

Bitterling Fish Optimizer (BFO)

BFO is employed to perform optimal feature selection, thereby improving the model's accuracy and efficiency in both information retrieval and the prediction of mental health outcomes. The proposed BFO enhances AI-driven knowledge retrieval and mental health prediction within digital libraries. Numerous species in nature employ specialised strategies to improve survival and reproductive success. The BFO algorithm draws inspiration from the unique reproductive behaviour of bitterling fish. In the wild, male bitterlings locate and defend freshwater mollusk shells, such as oysters, which serve as optimal breeding sites. Competition arises as larger shells offer superior protection for eggs. During this process, males display adaptive behaviours, including territorial defence and colour changes, to attract females and secure optimal spawning locations. The BFO algorithm replicates these competitive and intelligent selection behaviours to identify the most relevant features. Feature selection using BFO aims to maximise correlations among information retrieval patterns, cognitive load indicators, and emotional responses. By simulating natural selection and competition among bitterling fish, BFO ensures that the most informative attributes are accurately identified. The algorithm proceeds through the following stages:

Step 1: Setting Up

The initial positions of the bitterling fish within the search space are randomly assigned. Let C denote the number of decision-making factors, and M represent the population size. The position of the j-th fish, expressed as a vector, is defined in Equation (12).

$$W_j = [w_{j1}, w_{j2}, \ldots\ldots, w_{jC}], j = 1,2, \ldots\ldots, M \quad (12)$$

Where W_j is the position vector, w_{j1}, w_{j2}, ….., w_{jC} is the individual feature, C is the total number of choice vectors, and M is the total number of BFO.

Step 2: Function of the Objective

The issue to be addressed is represented by the objective function f (W_j). The goal is to either maximise or minimise the value of this function. The fitness of each fish is evaluated using Equation (13).

$$e(W_j) = objective\ Function\ to\ be\ optimized \quad (13)$$

Here, $e(W_j)$ denotes the fitness value, and the objective function represents the mathematical criterion used to assess the quality of a solution.

Step 3: The Movement of Bitterling Fish

Each bitterling fish navigates the search space by balancing exploitation and exploration. The movement is formalised in Equation (14).

$$W_j^{s+1} = W_j^s + \alpha_j.(W_{best} - W_j^s) + \beta_j.(W_j^s - W_{rand}) \quad (14)$$

Where the fish's location at time step is represented by W_j^s. The best fish (leader) is located at W_{best}. Here, W_{rand} indicates the location of a fish chosen at random, a_j is the attraction coefficient, and β_j is the repulsion coefficient.

Step 4: Assessment

The objective function evaluates the fitness of each fish at its updated position, providing a measure of how effectively the current location satisfies the optimisation criteria.

Step 5: Revise

The position of each fish is adjusted based on its fitness level. The new position of fish j is accepted if it results in an improved fitness value. The update procedure is defined in Equation (15).

$$W_j^{s+1} = \begin{cases} W_j^s\ if\ e(W_j^{s+1}) \leq e(W_j^s) \\ W_j^{s+1}\ \ if\ the\ new\ position\ is\ better \end{cases} \quad (15)$$

Step 6: Termination Standards

Termination criteria for the algorithm are defined to determine when the process should stop. Equation (16) presents the mathematical formulation of this termination condition.

$$Terminate\ if: s \geq S\ or\ {}_j^s max \quad (16)$$

Here, S represents the maximum population size.

Step 7: Analysis of the Results

Equation (17) represents the optimal solution obtained upon the termination of the process.

$$W_{best} = \arg\min_j e(W_j) \quad (17)$$

Where W_{best} is the best fish's ideal location when the algorithm is finished, $e(W_j)$ is the error function, signifying the problem's optimal solution. This approach effectively enhances mental health assessment and intelligent information retrieval within digital libraries. The BFO framework for AI-based user behaviour prediction is illustrated in Figure 6.

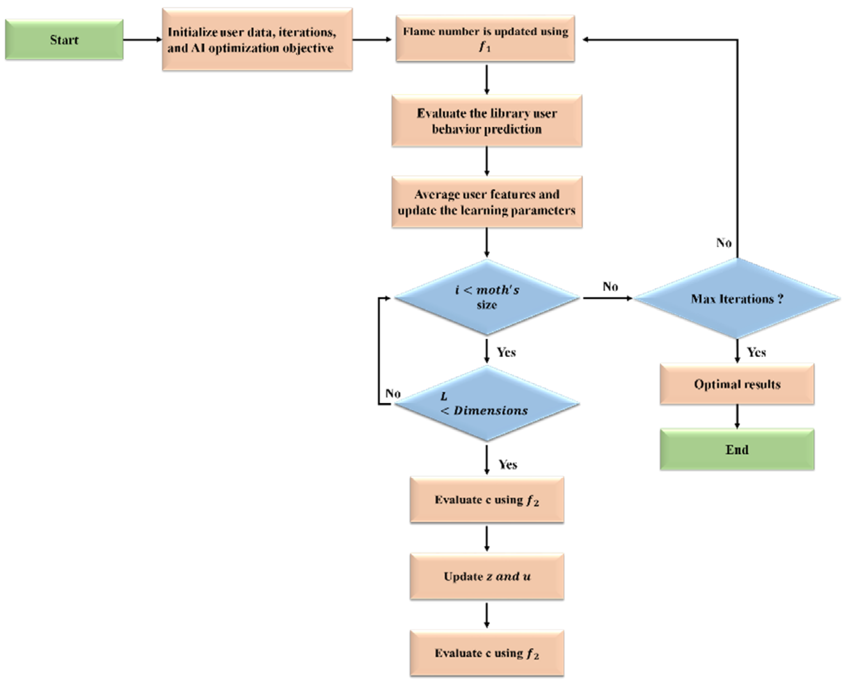

Figure 6: Flowchart of the BFO Algorithm for Library User Behavior Prediction.

Results and Discussion

Dataset

The Library Behaviour and Mental Health Dataset, available on Kaggle, contains user interaction logs from academic library systems, encompassing search queries, access frequencies, and feedback entries. It is designed to support analysis of information retrieval behaviours, digital literacy, and related mental health indicators, enabling the investigation of patterns linking library use with cognitive and emotional outcomes (Kaggle: https://www.kaggle.com/datasets/programmer3/library-behavior-and-mental-health-dataset). The proposed BFO-DFT algorithm markedly enhances personalised course recommendations, learning relevance, and user satisfaction compared to conventional approaches, as demonstrated by the highly effective results obtained from the Python-based implementation. These findings confirm the robustness of the model and its adaptability across diverse academic datasets.

Performance Evaluation

This 3D scatter plot illustrates the relationship between search queries, access frequency, and the User Engagement Score (Z value) derived from the Library Behaviour and Mental Health Dataset. The BFO-DFT model effectively identifies patterns indicative of mental health status and digital literacy. The 3D correlation of search queries, access frequency, and engagement level is presented in Figure 7.

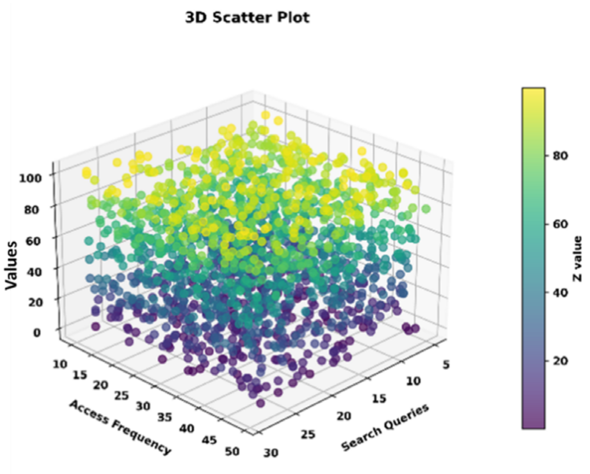

Figure 7: 3D Visualization of User Interaction Patterns in Digital Libraries using BFO–DFT Model.

This pair plot depicts the correlations among search queries, access frequency, and feedback sentiment within the Library Behaviour dataset. The distribution patterns highlight differences in user engagement and emotional responses. By applying the BFO model for well-being, correlations between information retrieval behaviours and indicators of mental health were effectively identified. The pairwise visualisation of digital literacy and sentiment patterns is shown in Figure 8.

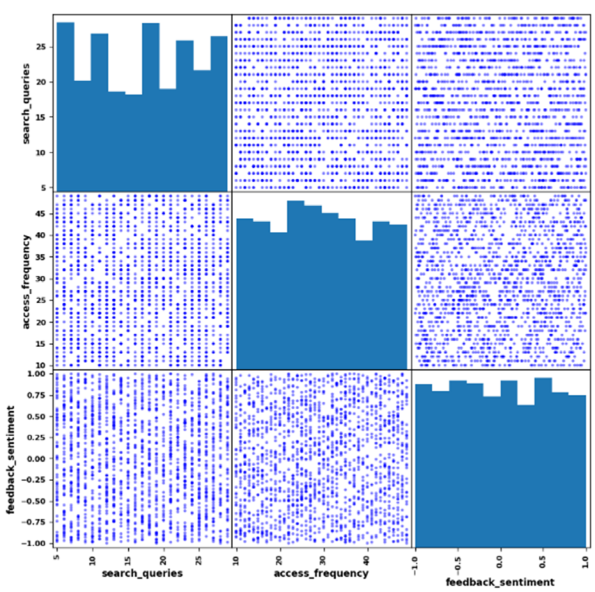

Figure 8: Correlation Analysis of Library User Interactions using BFO–DFT.

The 3D wireframe diagram illustrates the complex relationships among search queries, access frequency, and the Engagement Surface Value within library behaviour. The surface structure reflects variations in information-seeking behaviour and digital literacy, emphasising nonlinear fluctuations in user engagement intensity. The BFO-DFT model was employed to analyse intricate associations between information retrieval patterns, cognitive engagement, and mental health indicators, supporting intelligent and adaptive library management processes. Figure 9 presents a wireframe visualisation of user interaction dynamics within the library.

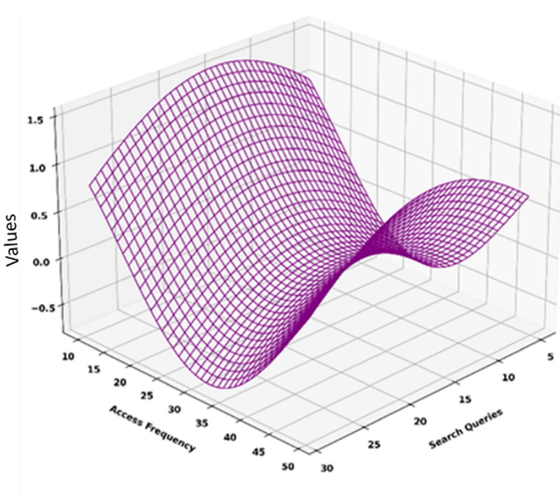

Figure 9: 3D Engagement Surface Analysis using BFO–DFT Model.

The Feedback Sentiment Distribution graph displays user sentiment scores ranging from -1.0 (negative) to +1.0 (positive), with most counts falling between 120 and 150 per sentiment score interval. The balanced sentiment assessment achieved using the BFO-DFT approach facilitated accurate prediction of user behaviour and enhanced recommendation performance. Figure 10 illustrates the Feedback Sentiment Analysis conducted with the BFO-DFT model.

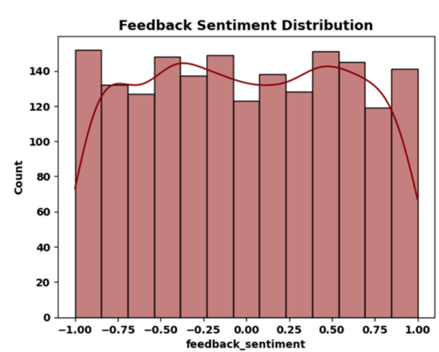

Figure 10: Sentiment Score Distribution for Library User Feedback.

Figure 11(a) depicts the density overlap between feedback sentiment and information retrieval efficiency, showing a moderate correlation of approximately 0.6, suggesting that the influence of sentiment on retrieval accuracy is balanced. Figure 11(b) presents a hexbin comparison of access frequency against retrieval efficiency, highlighting dense clusters (values = -7 to 8) that represent typical access patterns supporting retrieval performance. The BFO-DFT model was employed to analyse these correlations.

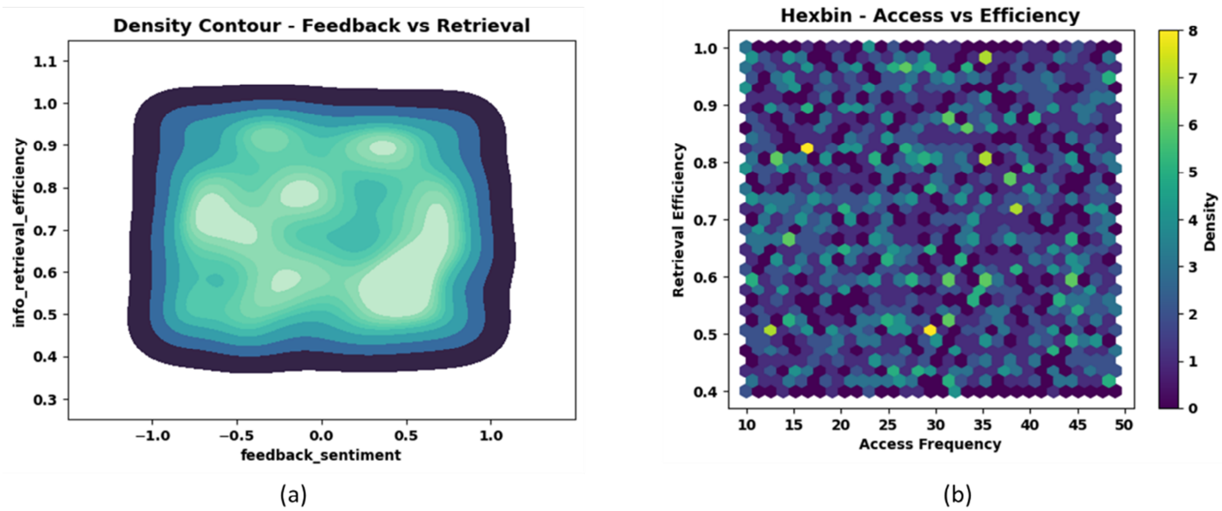

Figure 11: Visual Representation of (a) Density Contour of Feedback Sentiment vs Information Retrieval Efficiency (b) Hexbin Plot of Access Frequency vs Retrieval Efficiency.

The parallel coordinate figure illustrates the relationship among search queries, access frequency, and feedback sentiment across varying stress levels (0, 1, 2). Higher stress levels (represented by red lines) correspond to increased search activity and access frequency, accompanied by reduced feedback sentiment. This analysis reveals behavioural correlations that influence the accuracy of mental state prediction using the BFO-DFT model. The parallel coordinates visualisation of stress levels using the BFO-DFT method is presented in Figure 12.

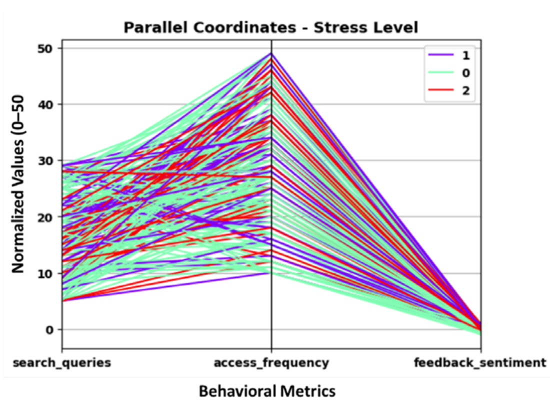

Figure 12: Stress Level Behavior Analysis using BFO-DFT.

Comparative Analysis

Compared with existing deep learning models, the discussion highlights that the proposed BFO-DFT approach significantly enhances both the efficiency of information retrieval and the accuracy of predictions. The performance of the CBF-CF method (Wayesa et al., 2023) is highly dependent on the quality and semantic precision of the data; insufficient book descriptions or sparse user profiles limit its ability to provide personalised recommendations for new users. Potential limitations of the machine learning techniques employed in the CATALYST platform (Ma, 2024) include restricted dataset diversity, algorithmic bias, and reduced adaptability to evolving user requirements within large-scale university library environments.

Precision: Measures the proportion of predicted positive outcomes that are correctly identified.
Recall: Represents the fraction of true positive cases that are successfully detected.
F1-Score: Defined as the harmonic mean of precision and recall, it provides a balanced assessment of overall performance between these two metrics.

The comparison indicates that the proposed BFO-DFT model (Precision = 0.83, Recall = 0.87, F1

= 0.85) outperforms both CBF-CF and CATALYST in terms of precision, reliability, and generalisation capability. Table 2 and Figure 13 present the comparative performance analysis between existing models and the proposed BFO-DFT method.

Table 2: Comparative Performance Analysis of the Proposed BFO-DFT Model and Existing Models (Precision, Recall, F1-Score).

Model	Precision (%)	Recall (%)	F1-Score (%)
CBF-CF (Wayesa et al., 2023)	0.63	0.40	0.52
CATALYST (Ma, 2021)	0.80	0.85	0.82
BFO-DFT [Proposed]	**0.83**	**0.87**	**0.85**

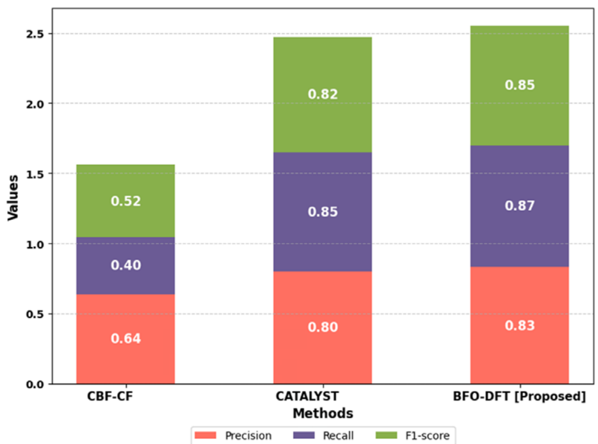

Figure 13: Evaluation of Course Recommendation Model Performance Using F1-Score, Precision, and Recall.

The proposed BFO-DFT model enhances personalisation, scalability, and predictive accuracy within digital library systems by addressing these limitations through intelligent feature selection, adaptive learning mechanisms, and improved analysis of data correlations.

Conclusion

The proposed approach has improved knowledge recall and the accuracy of mental health prediction by optimising both feature selection and model performance. The BFO-DFT model, which employs AI for prediction and optimisation, represents a significant advancement as it enhances knowledge management and digital literacy within academic libraries. The model effectively evaluates user behaviour, cognitive load, and emotional states through the combined application of feature selection using BFO and classification via Decision Tree and Random Forest techniques. TF-IDF and PCC are utilised to enable precise feature extraction and correlation analysis. It is noteworthy that small-scale testing of the BFO-DFT results is necessary to support the development of intelligent, adaptable, and user-friendly digital library ecosystems, by improving information retrieval accuracy, forecasting user satisfaction, and identifying stress indicators. In predicting user behaviour and satisfaction, the BFO-DFT model achieved high generalisation and reliable performance, with a precision of 0.83, recall of 0.87, and F1-score of 0.85.

Limitations and Future Scope

Ensuring interpretability and fairness in user-specific recommendations can mitigate bias and enhance transparency, which are often compromised in deep learning-based retrieval and recommendation models, particularly when data diversity and labeling quality are limited. Future work will incorporate transformer-based embeddings, multimodal data integration, real-time emotional analysis, and privacy-preserving federated learning to further enhance the model's retrieval accuracy, adaptability, and ethical implementation within library environments.

Project Support: The Supply-Demand Matching Employment Education Project of the Student Affairs Department of Chinese Ministry of Education: Research on the Management Path of Student Psychological Education Model in the Context of Internet (Project Number: 2025041433202)

References

Aldoseri, A., Al-Khalifa, K. N. and Hamouda, A. M. (2024). AI-Powered Innovation in Digital Transformation: Key Pillars and Industry Impact. *Sustainability*, 16(5): 1790. https://doi.org/10.3390/su16051790

Alon, N., Perret, S. and Torous, J. (2023). Working towards a ready to implement digital literacy program. *mHealth*, 9: 32. https://doi.org/10.21037/mhealth-23-13

Amin, S., Uddin, M. I., Alarood, A. A., Mashwani, W. K., Alzahrani, A. and Alzahrani, A. O.

(2023). Smart E-Learning Framework for Personalized Adaptive Learning and Sequential Path Recommendations Using Reinforcement Learning. *IEEE Access*, 11: 89769-89790. https://doi.org/10.1109/ACCESS.2023.3305584

Ayman, S. E. and Abo El Rejal, A. (2023). Ontology and Machine Learning-Based Recommender System for Teacher Resource Personalization. *Education Science and Management*, 1(3): 145-157. https://doi.org/10.56578/esm010303

Balcombe, L. and De Leo, D. (2022). Human-Computer Interaction in Digital Mental Health. *Informatics*, 9(1): 14. https://doi.org/10.3390/informatics9010014

Barman, B. (2024). Artificial Intelligence and Machine Learning in Libraries Transforming Information Access and Management. *RGU Journal of Social Science and Research*, 1(1): 1-7. https://doi.org/10.3254/rgujssr.v1i1.4

Camacho, H., Coto, M. and Jensen, S. P. (2021). Participatory methods to support knowledge management systems design in educational environments. *International Journal of Knowledge Management Studies*, 12(1): 34-54. https://doi.org/10.1504/IJKMS.2021.112210

Ezaldeen, H., Bisoy, S. K., Misra, R. and Alatrash, R. (2022). Semantics-Aware Context-Based Learner Modelling Using Normalized PSO for Personalized E-Learning. *Journal of Web Engineering*, 21(4): 1187-1224. https://doi.org/10.13052/jwe1540-9589.2148

Ikwuanusi, U. F., Adepoju, P. A. and Odionu, C. S. (2023). AI-driven solutions for personalized knowledge dissemination and inclusive library user experiences. *International Journal of Engineering Research Updates*, 4(2): 52-62. https://doi.org/10.53430/ijeru.2023.4.2.0023

Kanwal, M., Khan, M. A., Ismat, N., Khan, N. A. and Khan, A. A. (2024a). Machine Learning Approach to Classification of Online Users by Exploiting Information Seeking Behavior. *IEEE Access*, 12: 53234-53249. https://doi.org/10.1109/ACCESS.2024.3383444

Kanwal, M., Khan, N. A. and Khan, A. A. (2024b). A Machine Learning Approach to User Profiling for Data Annotation of Online Behavior. *Computers, Materials and Continua*, 78(2): 2419-2440. https://doi.org/10.32604/cmc.2024.047223

Ma, C. (2024). Implementing a Machine Learning-Based Library Information Management System: A CATALYST-Based Framework Integration. *International Journal of Advanced Computer Science and Applications*, 15(10): 605-612. https://doi.org/10.14569/IJACSA.2024.0151062

Ma, Y., Ouyang, R., Long, X., Gao, Z., Lai, T. and Fan, C. (2023). DORIS: Personalized course recommendation system based on deep learning. *PloS One*, 18(6): e0284687. https://doi.org/10.1371/journal.pone.0284687

Malekloo, A., Ozer, E., AlHamaydeh, M. and Girolami, M. (2022). Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights. *Structural Health Monitoring*, 21(4): 1906-1955. https://doi.org/10.1177/14759217211036880

Nasir, M. (2024). Use of AI Tools in Library Operation and Services. *The Seybold Report*, 19(5): 60-67. https://doi.org/10.5281/zenodo.11118327

Poquet, O. and de Laat, M. (2021). Developing capabilities: Lifelong learning in the age of AI. *British Journal of Educational Technology*, 52(4): 1695-1708. https://doi.org/10.1111/bjet.13123

Qiu, F., Zhu, L., Zhang, G., Sheng, X., Ye, M., Xiang, Q. *et al.* (2022). E-Learning Performance Prediction: Mining the Feature Space of Effective Learning Behavior. *Entropy*, 24(5): 722. https://doi.org/10.3390/e24050722

Rafi, M., Islam, A. Y. M. A., Ahmad, K. and Zheng, J. M. (2022). Digital Resources Integration and Performance Evaluation Under the Knowledge Management Model in Academic Libraries. *Libri*, 72(2): 123-140. https://doi.org/10.1515/libri-2021-0056

Rohayani, H., Rico, Armiwita, A. and Ermaini, E. (2024). AI-based Digital Literacy for Vocational High School Principals in Muaro Jambi Regency. *Proletarian : Community Service Development Journal*, 2(2): 80-86. https://doi.org/10.61098/proletariancomdev.v2i2.225

Romero-Ochoa, M.-A., Romero-González, J.-A., Perez-Soltero, A., Terven, J., García-Ramírez, T., Córdova-Esparza, D.-M. *et al.* (2025). Knowledge Management Strategies Supported by ICT for the Improvement of Teaching Practice: A Systematic Review. *Information*, 16(5): 414. https://doi.org/10.3390/info16050414

Samia, M. and Hind, B. (2024). Navigating Critical Thinking in the Digital Era: an Informative Exploration. *International Journal of Linguistics, Literature and Translation*, 7(1): 137-143. https://doi.org/10.32996/ijllt.2024.7.1.11x

Shahbazi, Z. and Byun, Y.-C. (2022). Agent-Based Recommendation in E-Learning Environment Using Knowledge Discovery and Machine Learning Approaches. *Mathematics*, 10(7): 1192. https://doi.org/10.3390/math10071192

Tahir, S., Hafeez, Y., Abbas, M. A., Nawaz, A. and Hamid, B. (2022). Smart Learning Objects Retrieval for E-Learning with Contextual Recommendation based on Collaborative Filtering. *Education and Information Technologies*, 27(6): 8631-8668. https://doi.org/10.1007/s10639-022-10966-0

Thaher, T., Saheb, M., Turabieh, H. and Chantar, H. (2021). Intelligent Detection of False Information in Arabic Tweets Utilizing Hybrid Harris Hawks Based Feature Selection and Machine Learning Models. *Symmetry*, 13(4): 556. https://doi.org/10.3390/sym13040556

Wajid, M. A., Zafar, A. and Wajid, M. S. (2024). A deep learning approach for image and text classification using neutrosophy. *International Journal of Information Technology*, 16(2): 853-859. https://doi.org/10.1007/s41870-023-01529-8

Wang, L., Zhang, J., Yang, H., Chen, Z.-Y., Tang, J., Zhang, Z. *et al.* (2025). User Behavior Simulation with Large Language Model-based Agents. *ACM Transactions on Information Systems*, 43(2): 1-37. https://doi.org/10.1145/3708985

Wayesa, F., Leranso, M., Asefa, G. and Kedir, A. (2023). Pattern-based hybrid book recommendation system using semantic relationships. *Scientific Reports*, 13(1): 3693. https://doi.org/10.1038/s41598-023-30987-0

Weck, M., Humala, I., Tamminen, P. and Ferreira, F. A. F. (2021). Knowledge management visualisation in regional innovation system collaborative decision-making. *Management Decision*, 60(4): 1017-1038. https://doi.org/10.1108/MD-01-2021-0064

Zhang, C., Zhang, S., Liu, Y. and Cheng, R. a. (2023). Star Catalog Based on Triangulation Networks, Angular Distance Weights, and Approximate Longitude Models. *IEEE Sensors Journal*, 23(19): 23597-23607. https://doi.org/10.1109/JSEN.2023.3306928

Chengzhe Guo Postdoctoral Fellow and Assistant Researcher at the Faculty of Education, Beijing Normal University. He also serves as Director of the Education Department at Shaoxing Jiangda Education Training School, and as a master's and doctoral supervisor at Krirk University and Lampang Inter-Tech College in Thailand. He holds degrees from Zhejiang University, Dhurakij Pundit University, and Krirk University. His research interests include educational psychology, educational management, and AI-enabled education. In recent years, he has published multiple papers in SCI and SSCI journals and has led several key national and industry-funded research projects in China.

Shangjie Meng is a Postdoctoral Researcher at Chongqing University of Posts and Telecommunications. He previously served in academic and administrative roles at Krirk University and Beijing Language and Culture University Bangkok College. His research focuses on educational management, educational technology, AI-enhanced learning, and smart campus development. He has participated in multiple Sino-Thai collaborative projects and published studies on digital pedagogy, learning analytics, and technology-driven educational innovation.