A Study on Data Analysis of Libraries in United States Using Artificial Intelligence

Dr.Yashodamma G. Selection Grade Librarian
Maharani women's Arts Commerce and Management College-Bangalore 560001
yashodaphd.2020@gmail.com

Abstract

This study investigates borrowing patterns, usage behavior, and circulation dynamics in U.S. libraries by applying Artificial Intelligence (AI)–driven analytical methods. Using machine learning models, clustering techniques, and regression analysis, the research identifies key factors influencing book popularity and borrowing frequency. Visualizations such as boxplots, scatter plots, correlation heatmaps, and predictive modeling provide detailed insights into how variables like event count, publication year, genre, and purchase count relate to book circulation. Results indicate that event_count is the strongest determinant of borrow_count, while publication year and purchase patterns contribute minimally. Classification models, particularly Logistic Regression, demonstrate exceptionally high accuracy (99%) in predicting book popularity, outperforming Decision Trees. K-Means clustering reveals three distinct usage groups—high, medium, and low circulation clusters—highlighting clear disparities in borrowing behavior. Overall, the study emphasizes the critical role of AI in understanding library usage, enhancing decision-making, and supporting data-driven management of library collections.

Keywords : Artificial Intelligence, Library Data Analysis,Borrowing Behavior Prediction Machine Learning Models, Book Popularity Classification, K-Means Clustering

1. Introduction

Artificial intelligence (AI) has rapidly evolved from a specialized technological innovation to a transformative force influencing nearly every sector, including higher education. As universities increasingly integrate AI-driven systems into teaching, research, and administrative functions, the ability to understand and interact with these technologies has become an essential competency. Academic libraries, long regarded as centers of knowledge access and information stewardship, now occupy a pivotal position in this technological transition. They not only rely on AI-enabled tools for cataloging, discovery, instruction, and user support, but also serve as critical environments where students, faculty, and researchers develop the skills needed to navigate an AI-driven world. This expanding role highlights the growing importance of AI literacy among academic library employees.

Despite the accelerating adoption of AI across higher education, the current level of AI literacy among library professionals remains uneven. Existing studies suggest that while librarians acknowledge AI's potential benefits, many lack a comprehensive understanding of AI fundamentals, practical applications, and ethical implications. Furthermore, emerging technologies such as generative AI introduce new complexities related to accuracy, privacy, intellectual property, and bias—issues that directly affect libraries' longstanding commitments to equity, intellectual freedom, and responsible information practices. These challenges underscore the need to assess how prepared academic library employees are to evaluate, adopt, and teach with AI tools in their daily work.

While several frameworks and conceptual models, such as algorithmic literacy and digital literacy, provide partial insight into the skills required for interacting with AI, a clear and consistent understanding of AI literacy specific to academic libraries is still underdeveloped. Existing literature points to a lack of validated measurement tools, limited professional development opportunities, and inconsistent integration of AI literacy into library training programs. As AI becomes increasingly embedded in research workflows and learning environments, bridging these knowledge gaps becomes essential both for effective library operations and for supporting users who depend on library expertise.

In response to these emerging needs, the present study investigates the current state of AI literacy among academic library employees. It aims to assess employees' understanding of AI concepts, identify gaps in their knowledge and confidence, and explore perceptions regarding the integration of generative AI tools in library services. By examining these factors through a structured and comprehensive approach, this research seeks to inform professional development initiatives, guide policy decisions, and support responsible AI adoption in academic libraries. Ultimately, strengthening AI literacy within the library workforce is critical to ensuring that libraries continue to uphold their mission as adaptive, ethical, and forward- looking institutions in the evolving landscape of higher education[1].

However, despite the growing presence of AI in other sectors, many academic libraries continue to face significant challenges in understanding, adopting, and effectively integrating these technologies into everyday operations. Librarians frequently report concerns related to job displacement, ethical implications, data privacy, and the reliability of AI-generated outputs—challenges that hinder confident and informed adoption.

Although interest in AI is increasing across higher education, a critical gap persists in the literature regarding the actual readiness and AI literacy of academic library employees. Existing research largely examines perceptions of AI or explores technical implementations, but far fewer studies assess the foundational knowledge, skills, and competencies that library staff need to use AI responsibly and effectively. Moreover, while institutions in technologically advanced regions are beginning to experiment with AI-enabled services, many libraries— particularly in developing contexts— struggle with limited technical infrastructure, insufficient training opportunities, and a lack of standardized frameworks for evaluating AI literacy. This disconnect between the promise of AI tools and the preparedness of library personnel creates a substantial barrier to meaningful, ethical, and sustainable integration[2].

The purpose of this study is to address this gap by examining the current state of AI literacy among academic library employees and identifying the specific areas where additional training, support, or policy development is needed. By assessing staff knowledge, confidence levels, and perceptions toward emerging AI tools—including generative AI—this research aims to provide an evidence-based foundation for developing targeted professional development programs and informed institutional strategies. Ultimately, enhancing AI literacy in academic libraries is essential not only for improving service delivery but also for ensuring that libraries remain proactive, ethical, and adaptable in an increasingly AI-driven information landscape[3][4].

2. Literature Survey

The paper "Research on Artificial Intelligence in Libraries" presents a comprehensive bibliometric analysis of AI-related library research from 2015 to 2024 and reviews how AI technologies are being integrated into various library services. It highlights a significant growth in publications—especially in medical and computer science domains—with the United States, China, England, and Canada leading scholarly output. The study discusses practical AI applications in libraries, including information retrieval, reference services, cataloging, education, and smart recommendation systems, while also addressing challenges such as technical limitations, ethical concerns, job displacement fears, and uneven adoption across developing countries. The paper concludes that although AI offers major opportunities to enhance library efficiency and user experience, librarians need improved training, policy support, and awareness to responsibly implement AI tools and uphold academic integrity.

As systematic reviews continue to grow in volume and importance—particularly in evidence- based medicine—the demand for faster, more efficient search strategies has increased. AI tools such as ChatGPT and other large language models (LLMs) are being explored as potential solutions to address time constraints, repetitive work, and search complexity. The paper evaluates the performance, advantages, and shortcomings of AI tools compared with human information-specialist expertise.

The study highlights that although AI tools can generate search strings and suggest synonyms or keywords at high speed, their outputs often lack the precision, reproducibility, and methodological transparency required for systematic review standards. AI-generated searches frequently miss essential controlled vocabulary terms (such as MeSH), produce logically inconsistent or overly broad search strategies, or introduce hallucinated information that undermines search accuracy. AI tools also lack domain awareness of database structures, indexing rules, and nuanced decision-making that trained librarians routinely apply when designing comprehensive search protocols. As a result, the authors argue that AI-generated search strategies, when used independently, risk compromising the completeness and reliability of systematic evidence retrieval.

The paper emphasizes that expert librarians bring essential competencies—such as critical appraisal of terminology, understanding of database architecture, iterative refinement of search syntax, and the ability to adapt strategies to diverse platforms—that AI tools have not yet replicated. AI also lacks accountability and cannot assess the methodological implications of search decisions, further limiting its suitability as a standalone solution. However, the authors acknowledge that AI tools may still serve as valuable assistants, capable of performing supportive tasks such as brainstorming keywords, improving efficiency during the early exploratory search phase, and automating aspects of documentation or deduplication.

Ultimately, the paper concludes that while AI tools show promise as supplementary aids, they cannot replace librarians in the systematic search process. Human oversight remains essential to ensure accuracy, transparency, reproducibility, and adherence to established systematic review standards. The authors advocate for a collaborative model in which librarians integrate AI tools thoughtfully and cautiously into their workflows while maintaining professional judgment and methodological control. The paper calls for continued research on AI–human collaboration, development of validation frameworks, and establishment of ethical guidelines for AI use in evidence synthesis.

This paper introduces the design and implementation of an intelligent book recommendation system specifically tailored for university contexts. This proposed system utilizes deep learning models embedded within a hybrid recommendation framework. The overarching goal is to address the shortcomings of traditional university library recommendation systems, which primarily rely on conventional collaborative filtering and keyword-based content similarity. These traditional methods are often vulnerable to cold-start issues, data sparsity, and weak semantic comprehension.

System Architecture and Components

The system employs a hybrid approach that merges content semantics with collaborative behavior.

1. Natural Language Processing (NLP) / BERT: State-of-the-art NLP techniques are utilized for the semantic extraction of book metadata, including the title, abstract, and subject tags. The model uses BERT (Bidirectional Encoder Representations from Transformers) for textual semantic comprehension.

o BERT Mechanism: BERT is a deep contextual language model based on the Transformer architecture. Its primary contribution is its bidirectional training mechanism, which allows it to learn the context of a word from both the left and right sides via multi-head self-attention layers. BERT is trained using two unsupervised objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).

o Content Embedding: BERT embeds textual inputs (titles, summaries, user reviews) to generate semantic-rich embeddings, typically derived from the last hidden state of the token.

2. Deep Autoencoder: A deep autoencoder is used to capture latent user-book interactions derived from borrowing histories.

o Autoencoder Mechanism: An autoencoder is a neural network trained to acquire efficient compressed data representations in an unsupervised manner. It consists of an encoder (reducing input to a latent space vector, $h = f(\mathbf{x})$) and a decoder (reconstructing the original input, $\mathbf{x}' = g(h)$), minimizing the reconstruction error $\mathcal{L}(\mathbf{x}, \mathbf{x}')$.

o Feature Learning: Autoencoders are beneficial in recommendation systems for compressing high-dimensional user-item matrices or learning latent features from structured data, effectively mapping sparse data into a dense vector space by extracting nonlinear dependencies.

3. Hybrid Integration: The proposed hybrid model overcomes weaknesses like cold-start issues and data sparsity by combining the proposed model's generated content representations (from BERT) with user preference vectors (learned through autoencoders). These two feature vectors are concatenated via a joint representation layer and fed into

a downstream neural network to predict user-item interactions.

Dataset and Implementation

The research used a real-world dataset compiled from thousands of books and borrowing records to evaluate and train the model. The training and evaluation employed a merged dataset based on the shared ISBN field from the Book-Crossing dataset (for collaborative filtering data) and the Goodreads Books dataset (for content-based features like descriptions, genres, and ratings). The dataset was divided into an 80:20 train-test split ratio.

Data preprocessing involved handling missing values, eliminating duplicates, and normalizing textual fields. Text data for BERT was tokenized using Word Piece tokenization, and numerical features were normalized to a 0–1 range using Min-Max scaling. The system was implemented using Python 3.12 with libraries such as matplotlib, numpy, and scikit-learn, running on an Intel i7 system with an NVIDIA Tesla GPU.

Experimental Outcomes

The performance of the proposed hybrid model was significantly better than traditional recommendation baselines, demonstrating the ability to provide reliable and targeted recommendations.

Algorithm / Performance Metrics	Accuracy	Precision	Recall	F1 Score
RNN	92.4%	92.5%	92.5%	92.8%
BiLSTM	94.5%	94.5%	94.5%	94.5%
BERT	94.8%	94.7%	94.7%	94.8%
Autoencoder	95.8%	95.8%	95.0%	95.8%
PROPOSED Hybrid Model	97.4%	96.8%	96.2%	97.0%

The high performance of the proposed model is attributed to its optimized architecture and the incorporation of innovative deep learning strategies that successfully capture complex data relationships. The ROC curve analysis confirmed the highest discriminative power and lowest false alarm rates for the proposed model. Deployment results showed significant increases in student engagement, search efficiency, and content discoverability in digital library portals.

Future research suggestions include incorporating user-provided data like social activity, investigating Variational Autoencoders (VAE) for better regularization, and adapting the model for real-time environments with feedback loops[5].

This research proposes an evaluation framework utilizing deep machine learning for building a digital library recommendation system based on Metadata for Arabic and English languages. Metadata, described as data about data, is essential for digital library integration, aiding in description, browsing, transfer, retrieval improvement, preservation, and interoperability.

Methodology and Data

The study focuses on using deep learning models combined with word embedding to achieve high-performance metadata classification and improve services in digital libraries. The datasets used were extracted from the Saudi Digital Library (SDL) database, which included 2447 items for the English dataset and 2098 items for the Arabic dataset, primarily focusing on computer science categories.

Data Processing Steps: Data cleaning involved converting text to lowercase and removing white space and punctuation. For Arabic text, specific mechanisms were used to remove unnecessary characteristics and diacritics. Tokenization (dividing text into individual words) was performed using the split() function, and stop words were removed using NLTK for both languages.

Similarity Metrics and Weighting: The system computes similarity between documents by combining weighted scores derived from metadata fields:

- Title similarity was given a weight of 3.5.
- Abstract similarity was given a weight of 4.0.
- Subject similarity was given a weight of 1.5.
- Journal similarity was given a weight of 1.

The similarity methods evaluated were:

1. TF-IDF (Term Frequency-Inverse Document Frequency) similarity: Used to compute semantic similarities and transform text into feature vectors, applied to the Abstract and Title.
2. Cosine similarity: A symmetrical algorithm used to find similarity between items.
3. Jaccard similarity (JS): Finds textual similarity at a character level, not based on meaning.
4. Semantic Similarity (Word2Vec): Based on the likeness of word meaning, crucial for Natural Language Processing (NLP).

Experimental Outcomes for Similarity

Similarity scores were compared using Top N Accuracy:

Dataset	Cosine Similarity	Jaccard Similarity	TD-FID Similarity	Semantic Similarity (Word2Vec)
SDL Arabic	30%	20%	20%	96%
SDL English	4%	55%	14%	84%

Semantic similarity based on Word2Vec yielded significantly better results (96% for Arabic and 84% for English) compared to the other methods. This high performance is due to its reliance on the meaning of words rather than their frequency or weight, allowing it to detect items related by similar meaning even if they use different words.

Metadata Classification Results

The study also tested the performance of traditional and deep learning classification models (SVM, CNN, RNN, LSTM) on Subject, Title, and Abstract metadata.

Model	Dataset	Classifier	Accuracy	Notes
SVM	Arabic	Title	91%	Better than English due to less class overlapping.
	English	Title	70%	
	Arabic	Abstract	93%	
	English	Abstract	73%	
CNN	English	Title	93.70%	Improved performance over SVM.
	Arabic	Title	88%	

	English	Abstract	90%	Improved performance over SVM.
	Arabic	Abstract	93%	Same as SVM.
RNN & LSTM	Arabic	Title (LSTM)	86%	RNN showed poor results; LSTM improved them.
	English	Title (LSTM)	44%	
	Arabic	Abstract (LSTM)	81%	
	English	Abstract (LSTM)	45%	RNN/LSTM performance is strong on titles but poor on full text.

In conclusion, the proposed methodology successfully utilizes metadata for digital library recommender systems, emphasizing the superior performance of semantic similarity methods like Word2Vec for personalized recommendations[6].

This document presents a demonstration proposal for DiffeRT, an open-source, Differentiable Ray Tracing (DRT) toolbox. DiffeRT is designed for Machine Learning (ML) and optimization applications within the field of radio propagation.

Background and Technology

Ray Tracing (RT) is a standard technique used across physical optics, computer graphics, and communication systems to model high-frequency wave propagation by simulating wavefronts as rays. The evolution to Differentiable Ray Tracing (DRT) leverages automatic differentiation (AD) to make ray paths and associated outputs differentiable with respect to system parameters, enabling seamless integration into gradient-based ML and optimization workflows.

DiffeRT is built on powerful foundational technologies:

1. JAX Framework: Used for efficient differentiable programming. JAX is an array programming library featuring just-in-time compilation, automatic differentiation (AD), and native scalability to GPUs and TPUs.
2. Equinox Library: Used as a flexible ML framework.

DiffeRT extends its predecessor, DiffeRT2d, to support 3D scenes with enhanced performance and higher-level electromagnetic (EM) field computation features. The core functionality involves computing gradients of ray path outputs with respect to path candidates, which is valuable for tasks such as optimizing propagation scenarios or training ML models for RT.

Comparison with Alternatives

The landscape of radio propagation DRT tools is nascent, with Sionna being the primary open- source alternative. While Sionna is a comprehensive link-level simulator spanning channel coding to coverage map simulation, DiffeRT differentiates itself by narrowly focusing exclusively on RT-specific functionalities.

DiffeRT offers several key advantages due to its architecture:

- Flexibility: Users can choose between computing all paths in a scene or tracing user- defined subsets of path candidates.
- Scalability: Built on JAX, it is natively scalable to GPUs and TPUs, ensuring high performance.
- Customizability: Differentiable outputs allow seamless integration into gradient-based ML and optimization workflows.

Furthermore, DiffeRT provides multiple visualization backends (Plotly, Matplotlib, and VisPy) for efficiently visualizing large scenes and creating interactive graphics. It also includes a built- in method to load scene files created for Sionna, promoting reuse of existing material.

Demonstration and Applications

The demonstration is structured to showcase DiffeRT's capabilities through an interactive session. Key segments include an overview of its JAX/Equinox architecture, live examples of 3D RT in complex scenes, and a demonstration of ML Integration where DiffeRT computes differentiable ray paths to train a simple neural network for optimizing ray path selection. A side-by-side comparison with Sionna is also planned to illustrate DiffeRT's focused functionality.

DiffeRT has already been utilized in research, such as training an ML model for sampling candidate ray paths and comparing different RT implementations. The demonstration highlights DiffeRT's potential to drive innovation in RT for ML and optimization applications[7].

This research focuses on the development of Angkaew, an open-source automated library system integrated with a book recommendation system, specifically tailored for small libraries (those having around 3000–10000 books and limited loan records).

System Framework and Methodology

The system is based on the existing open-source framework OpenBiblio. The Angkaew framework implemented improvements for Thai language support, web responsiveness, compatibility with PHP 7.0 or higher, and the core book recommendation system.

The recommendation method employs a combined strategy using the Support Vector Machine (SVM) model trained on multiple features. This approach aims to provide recommendations even without extensive historical loan data.

Features used for SVM Training:

1. Title Similarity: This is computed using the Damerau–Levenshtein distance algorithm. This method is suitable for Thai language string comparison and is robust against common misspellings such as the transposition of two characters. The algorithm calculates the minimum number of single-character edits (insertion, deletion, mismatch, or transposition) required to change one string into the other.
2. D.D.C. (Dewey Decimal Classification) Matching: The standard classification method used in Thai libraries. The method utilizes the ten first-level D.D.C. categories as the main concept.
3. Bibliographic Information Similarity: This feature combines multiple metadata elements using a weighted equation:

o Year of publication: Calculated as the absolute difference between publication years, based on the assumption that users prefer books published close to those they select.
o Topic (book category) and Author: Calculated as a binary match (0 for match, 1 for no match).
o Number of views: Computed based on the total views by users registered in the system.
o The combined equation: $\text{Bib data} = w_1|y_s - y_r| + w_2(T_m + A_m) + w_3(N)$, where $y_s - y_r$ is the difference in publication year, T_m is the topic match, A_m is the author match, and N is the number of views.

The framework uses the LibSVM Version 3.24 classifier. The system recommends a total of nine books to users (three books based on the highest scores from each of the three feature categories).

The experiment was conducted at the Banpasao Chiangmai school, collecting data on 2555 books in the database and 4612 books in the loan record. Participants (60 primary students and 55 high school students) rated their interest in the recommended books on a five-point scale.

The overall results showed that the best performance in terms of high positive evaluation ("Interested" + "Very interested") was achieved by the D.D.C. matching and Bibliographic Information feature set (32.2% + 17.4% = 49.6%). However, the study noted that the SVM combining all features showed the highest rate of "Already Bought or Read" (11.3%), indicating strong alignment with user interests.

The conclusion emphasized that the approach of combining title similarity and bibliographic data performed better than relying on a single feature alone. This method offers an effective solution for small libraries lacking the resources or extensive loan records necessary for complex machine learning models[8].

This study focuses on constructing a usage behavioral model for university library electronic resources by applying the Multilayer Perceptron (MLP) deep learning model to predict reader behavior intention and the quality of use.

Model Framework and Data

The research utilized a model framework that composites the Unified Theory of Acceptance and Use of Technology (UTAUT) with Website Service Quality (WSQ). The five main facets explored were Performance Expectancy (PE), Effort Expectancy (EE), Social Influence (SI), Facilitating Conditions (FC), and Website Service Quality (WSQ).

Data was collected via questionnaires using a 7-point Likert scale (ranging from highly satisfied/agree to highly dissatisfied/disagree). The study involved 1,071 valid questionnaires collected from public and private university students (fourth-grade college students and second-year master students) in Taiwan.

Methodology: Multilayer Perceptron (MLP)

The data collected was trained and tested using the Neural Network–Multilayer Perceptron (NN-MLP), a feedforward artificial neural network. The neural network excels at mapping input data to output data through arbitrary degrees of nonlinearity and demonstrates strong learning ability.

The MLP model was implemented using Keras modules in Python.

- Data Preparation: Raw questionnaire data underwent filtering and flattening procedures, converting features and label values into NPZ format.
- Architecture: The computing model includes an input layer, a hidden layer, and an output layer, with data divided into an 8:2 train-test split.
- Functions: The Keras Sequential model used the activation functions "Relu," "Adam," and "linear" to obtain output. The input layer transmits values to the hidden layer, which applies weighted accumulation and activation function conversion before transmission to the output layer.
- Optimization: The study experimented with adjusting the number of hidden layer neurons (from 3 to 36) and epochs (from 100 to 2000).

The MLP model demonstrated high predictive capability for usage behavior, achieving an accuracy of 90.75% or more. The input layer used 24 scale indexes representing the Web service quality dimension, and the output layer was defined as user behavior.

The optimal configuration achieved the best result with a difference (error) of 0.647 (relative to the 7-point scale), corresponding to an accuracy rate of 90.8%, using 6 neuron units and 500 epochs.

The study highlighted key insights regarding network design:

- Excessive neurons in the hidden layers can lead to overfitting, especially with limited data.
- Too many neurons in the hidden layer can also increase training time, potentially making the process infeasible.

The research affirmed that deep learning regression models, such as MLP, can be effectively applied to predict user intent and behavior, aiding decision-makers in enhancing library services (e.g., adding electronic multimedia resources or selecting future books)[9].

This research addresses the challenges of information retrieval in digital libraries (DLs) characterized by diverse, heterogeneous, and massive multimedia data. The study constructs a cross-media semantic retrieval framework based on Deep Learning (DL) to enhance DL interactivity and knowledge services.

The Need for Deep Learning in Retrieval

Traditional retrieval systems, which mainly rely on text annotation, simple keyword queries, or manually designed features, are inadequate for handling complex dimensional data and cannot meet users' deep needs for knowledge. DL technology is proposed because it can automatically extract multi-level features from datasets, find associations between different data modalities (text, image, sound, video), and efficiently process complex information.

The core challenge for cross-media retrieval is bridging the semantic gap—the difficulty in searching for semantic relationships across highly heterogeneous information. DL solves this by utilizing semantic mapping and similarity calculation to simplify the process and improve accuracy.

Deep Learning and Retrieval Requirements

Deep learning, first proposed by Hinton in 2006, involves complex hierarchical structures (deep structures) that simulate the human brain's signal processing to achieve feature extraction, transforming low-dimensional features into more abstract high-level features.

Cross-media retrieval in DLs requires:

1. Conversion Mechanism: Establishing a mechanism between different multimedia resources to reflect their potential relevance through unified identification.
2. Unified Processing: Conducting analysis and processing in a unified data description manner to break through the semantic gap.
3. Semantic Understanding: Mining semantic related keywords, aligning contextual representation with thematic information, and extracting features from different spaces to explore associations.

Cross-media Semantic Retrieval Framework

The proposed framework integrates DL into a three-layer architecture:

1. Technology Layer: Focuses on the collection and processing of resources.
o Resource Collection and Maintenance: Utilizes intelligent proxy software to automatically collect, capture, and recognize data from web pages, then standardizes and stores them in a multimedia resource database.
o Feature Extraction and Semantic Association: DL models like CNN (for images) and RNN (for text, capturing long-distance dependencies) are used to extract features and map low-level features to the semantic level, eliminating semantic gaps.
2. Semantic Layer: Focuses on knowledge representation and organization.
o Ontology Construction and Self-learning: DL helps in forming dynamic cross- media ontologies based on extracted features and semantic associations, transforming low-level features into high-level semantic spaces. This involves Media Feature Learning (extracting single modal features and semantic clustering centers) and Knowledge Representation Learning (inferring knowledge relationships between entities based on cross-modal correlation).
3. Retrieval Layer: Handles the user interaction and result presentation.
o Retrieval Algorithm Design: Converts multiple forms of feature information into a unified space through semantic mapping, obtaining global features and topic distribution (e.g., using Deep Convolutional Neural

Network) and calculating similarity (e.g., Euclidean or Manhattan distances).

Related Technologies

To achieve semantic matching, heterogeneous features must be mapped to the same space. Key technologies utilized include:

- Information Extraction (IE): Deep learning models perform well in automatically extracting specific entity, relationship, event, and fact information from natural language text.
- Implicit Feature Association: Establishing the mapping relationship between image and text feature vectors in an isomorphic space to discover hidden associations.
- Spatial Mapping Technology: Mapping different data types to a shared space, often using training classifiers to semantically classify hidden layer features to achieve the same dimension.
- Cross-media Correlation Computer Technology: Techniques like Canonical Correlation Analysis (CCA) and correlation calculation are used to project variables linearly and calculate correlation coefficients between different modal data. Other feature mapping methods include SLSA, MDBN, EMK, and MLP.

The goal of this framework is to provide users with more targeted retrieval services by establishing semantic associations between different resources[10].

This study investigates the application of Deep Learning (DL) technology to optimize and innovate Knowledge Discovery (KD) Services in digital libraries (DLs).

Theoretical Foundation

Knowledge Discovery (KD) is defined as the process of refining, processing, and organizing explicit and implicit information according to user needs to form knowledge products. It involves collecting user behavior data to predict potential knowledge needs.

Deep Learning (DL) technology is characterized by its deep structure, surface structure model, and superior learning, computing, and expression capabilities. Its deep network structure is similar to neural networks and enables hierarchical feature extraction and global feature recognition, helping to eliminate the semantic gap between different media information.

DL/KD Correlation: Traditional KD methods often lack analysis of inherent resource relationships or semantic verification capability. DL overcomes this by abstracting original data through a deep structure model, extracting feature attributes layer by layer, and utilizing unified semantic description logic to improve the efficiency of knowledge aggregation.

Role of Deep Learning in KD Service

DL promotes several improvements in digital library services:

1. Improving Digital Resource Integration Efficiency: DL uses deep network structures to discover relationships between isolated network resources (data islands) and concatenate them using unified semantic logic, providing convenience for resource integration and reducing KD costs.
2. Improving Cognitive Level of Knowledge Discrimination: DL structures, similar to the human brain, transform the knowledge system into a complex network structure, facilitating the analysis and cognition of multimodal data. This enhances personalized services by focusing on contextual differences and simulating individual knowledge memory structures.
3. Improving Service Function of Digital Resources: DL helps break information barriers, leading to the

establishment of a unified digital resource retrieval interface. DL enables the efficient preprocessing of resource data, transforming collected data into input metadata that meets requirements for subsequent data mining and knowledge aggregation.

Knowledge Discovery System Architecture

The proposed system architecture is layered to provide technical support for KD:

1. Basic Resource Layer: The foundation, containing the library resource repository, user information repository, and tacit knowledge repository. This layer uses subject classification, pattern evaluation, and DL/clustering analysis to classify and display digital resources.
2. Knowledge Discovery Layer (Core): Utilizes DL and data mining algorithms to uniformly process, analyze, semantically associate, and mine heterogeneous data. It enables self-learning, discovers required knowledge/rules, and predicts knowledge matching based on personalized user needs.
3. Knowledge Service Layer (Top): Provides personalized services like information retrieval and customization through a visual user interaction interface. It integrates semantic retrieval and information visualization technologies.

Innovative Knowledge Discovery Services

The research proposes several innovative DL-based services:

1. Integrated Knowledge Retrieval: Moving beyond single modal searches, this system constructs an integration of multimodal cross-media knowledge resources (text, speech, gesture, vision input). It provides content-consistent resources like books, videos, research reports, and patents related to a theme.
2. Panoramic Knowledge Navigation: Innovates traditional navigation (which is limited in scope and mode) by comprehensively processing cross-source and cross-mode knowledge resources to form an integrated knowledge network.
3. Contextualized Knowledge Recommendation: Embeds knowledge services into the user's learning or research task context, recommending integrated multimodal knowledge to support overall and multi-dimensional understanding.
4. Embedded Knowledge Consulting: Provides deep consulting services, especially for scientific research users, intelligently embedded throughout the research workflow (e.g., project application, execution, problem research).
5. Automated Knowledge Q&A: An effective way to query the rich knowledge stored in knowledge graphs. It extracts key information from the user's natural language question, infers from the knowledge graph, and feeds back answers. Future development includes Visual Question Answering (VQA), integrating image and voice input[11].

This project focuses on establishing a user-oriented personalized teaching resource recommendation system for digital libraries, leveraging deep neural network technology to combat "information overload" and "information trek" in online education.

System Design and Hardware

The system employs a three-layer architecture. The hardware design is optimized for data processing, utilizing the Conroe-i70K processor (3.0 GHz main frequency, eight cores, fourteen threads). This processor is intended to support various wireless services, including Bluetooth, universal package wireless services, and 5G technology.

The software design focuses on effective management of multimedia teaching resources (text and video) and construction of a database using standard data forms, such as "student personal data form" and "teacher education resource form".

Deep Mining Methodology

The core of the system relies on deep learning to mine data resources and capture user intent.

1. User Modeling: Deep learning technology is applied to create a user portrait model, reducing dependence on underlying data to quickly obtain massive deep information for better resource recommendation.

○ Data Types: User data includes basic information (name, age, gender, major), behavior information (records in the network environment), and context information (geographic location, device, weather).

2. Feature Extraction and Weighting: The system assigns differentiated weights to user portrait features based on their importance. It uses a dual-layer GRU (Gated Recurrent Unit) structure to extract both prolonged user interests (extended) and brief user interests (transient preferences).

3. Recommendation Model: The personalized content recommendation model (Figure 3 in the source) takes high-dimensional space vectors (mark vectors of the dataset and database) as input.

○ Calculation: The recommendation analyzes the correlation between user feature vectors (v) and resource vectors ($c*$) using the Cosine Distance method. The propagation mode of the relationship is governed by equations involving the update gate (F) and reset gate (Z). The model utilizes deep learning techniques to analyze the timing characteristics of users and integrates them with context.

The model was tested using the Amazon database.

Metric	Proposed Textual Method	Logistic Regression	Neural Network
Accuracy	77.34%	76.87%	76.57%
AUC	92.55% (in categorization)	91.47%	91.47%

The proposed algorithm demonstrated improved accuracy and superior AUC performance compared to conventional methods (logistic regression and neural networks), which is attributed to its deep learning approach in uncovering the internal relationships within user behaviors and fully considering timing characteristics.

Furthermore, the system showed high execution efficiency. Comparison of time consumption indicated that the textual method was competitive with or slightly faster than the logistic regression and neural network algorithms, especially when generating larger numbers of recommendations (25 results).

The findings confirm the viability of the innovative concept introduced, demonstrating enhanced precision and reliability in personalized recommendations[12].

This study proposes a prediction model for university library borrowing behavior, integrating Big Data and Machine Learning techniques, specifically using a Radial Basis Function (RBF) neural network optimized by the Ant Colony Algorithm (ACO). The goal is to improve the efficiency and accuracy of borrowing recommendations and estimations.

Big Data and Library Context

Public libraries aim for social and cultural benefits, emphasizing the dissemination ability of paper books, which necessitates digital transformation. The project focuses on big data analysis of user behavior to establish a heuristic borrowing process.

Data Resources of Public Libraries:

1. Digital Book Big Data: Electronic text conversions, machine translation versions, and audio books.
2. Book Lending Card Big Data: Traditional metadata (title, author, publisher) supplemented by word frequency feature codes and style identification codes for comprehensive query logic.

3. Member Big Data: Personal details and online/physical borrowing records.

Heuristic Recommendation Strategy: The system categorizes book proposals based on four priority levels: 1) Books with similar keywords/titles previously borrowed; 2) Books by the same or related authors; 3) Books aligned with the reader's occupation/work nature (e.g., management books for executives); and 4) Books related to a specific keyword/abstract.

The study introduces an innovative machine learning method to predict borrowing behavior accurately, avoiding defects in traditional methods (like the inability of basic card data to fully reflect book types) by leveraging complex algorithms.

Prediction Model Methodology

1. Feature Code Generation: The system digitizes book content information using Hanwang laser scanning and IFLYtek machine reading for word frequency analysis. Two convolutional neural networks (CNNs) generate independent data columns, converging into a 64-bit code comprising lexical feature code (based on word order data) and book-style code (based on original data). These codes are compared with the member's reading habit codes to determine a recommended order.
2. Radial Basis Function (RBF) Neural Network: The RBF neural network is used to solve the nonlinear high-dimensional problem of library borrowing. It transforms input information into a higher dimension via the hidden layer, turning it into a decomposition problem.
3. Ant Colony Algorithm (ACO) Optimization: ACO, a technique inspired by natural processes, is integrated to tune the RBF parameters (node center ϵ_i and spread δ_i) to achieve peak precision in predictions. ACO initializes the colony location and pheromone, using path selection rules based on pheromone concentration to ensure the ants converge to an optimal solution.

Confirmatory Experiments and Results

Using a university library as a sample, the model was tested for its ability to estimate loan volumes.

- Overall Prediction Accuracy: The model accurately estimated the total loan volume, with a deviation of less than 2.5% in the estimated monthly loan volume for the next two months.
- Prediction by Category: The model accurately estimated loan volume for various categories of books, with estimated errors generally remaining below 2.5% (e.g., Literature error: 1.03%, Medicine error: 0.09%, Culture error: 2.51%).
- Prediction by Reader Type: The model accurately predicted loan status based on gender and grade. Female students were observed to borrow significantly more books than male students, which is consistent with reality, and the estimation error for these groups was very small (e.g., female 2021 error: 1.81%; male 2021 error: 0.24%).

The application of complex data mining techniques (RBF and ACO) to library borrowing prediction represents an innovation over traditional statistical analysis or simple mathematical models, enhancing the comprehensiveness and foresight of predictions in the era of big data[13].

This research proposes the development of an innovative, open-source offline mobile application (app) designed to create a Smart Learning Environment for rural empowerment, particularly in areas like India where poverty, illiteracy, and poor network connectivity are challenges.

Motivation and Problem Addressed

Traditional libraries (public and rural) require individuals to learn primarily from physical books, a challenge for those needing guidance. Rural areas lack proper education systems and suffer from limited digital literacy and unreliable network access, which impedes online learning. The app aims to provide a guide for basic knowledge, allowing people to learn at their own pace and in the comfort of their homes.

Proposed System and Features

The proposed system is an interactive app for e-learning, which promotes engagement through multimedia information and is designed to work effectively offline.

Key Features:

- Offline Functionality: The main feature, allowing users to download and browse material without an internet connection, ensuring continuous education even with spotty connectivity or unreliable electricity supply.
- Interactive Content: Utilizes quizzes and games (like "missing letter" for English or addition/subtraction games for mathematics) to enhance learning interest and develop problem-solving skills. Game-based learning is seen as more effective than traditional E-learning platforms for emotional engagement and motivation.
- E-learning Resources: Includes Basic English (focusing on alphabet, grammar, and 850 Basic English words) and Basic Mathematics (arithmetic, shapes, formulas, fractions).
- Admin Interface: A central hub for administrators to manage users, configure settings, monitor quiz programs, and update content.

Technical Implementation

The system is developed using Django, HTML, CSS, and Bootstrap. Python IDLE is mentioned as the default IDE.

The design overcomes specific technical limitations related to offline use and data updates:

- Database Connection: The Django framework is used because it contains an in-built database, simplifying the connection and reducing application size.
- High Performance: Bootstrap is used to reduce code lines, making the application run faster.
- Offline Data Retrieval: The app runs without internet by using a local API key to fetch data in the database.
- Content Updates: To manage the difficulty of updating content offline, a cache of commonly used dynamic data is created. This allows the app to be updated frequently and efficiently (low MB usage) when internet access is briefly available.

The overall benefit is an enhanced, flexible, and accessible learning experience, particularly valuable for students with physical disabilities, allowing them to learn at their own pace in a suitable environment[14].

2.1 Problem Definition

Libraries generate massive volumes of data through borrowing records, user events, acquisitions, and genre distributions. However, these data are often underutilized due to traditional manual analysis methods that are insufficient for uncovering hidden patterns or predicting user behavior. The lack of automated, intelligent analytical frameworks limits the ability of libraries to understand circulation trends, identify high-impact resources, and anticipate borrowing patterns. Specifically:

- It is unclear which factors most strongly influence book popularity.
- Traditional statistical methods cannot accurately classify books as popular or non- popular.
- Libraries struggle to segment resources based on real user engagement.
- Predictive insights required for strategic decision-making (e.g., which books to acquire more of, which genres users prefer) are limited.

Therefore, the core problem addressed in this study is the absence of an AI-driven data analysis framework capable of modeling, predicting, and classifying borrowing behavior in library collections, enabling libraries to make evidence-based decisions and optimize services.

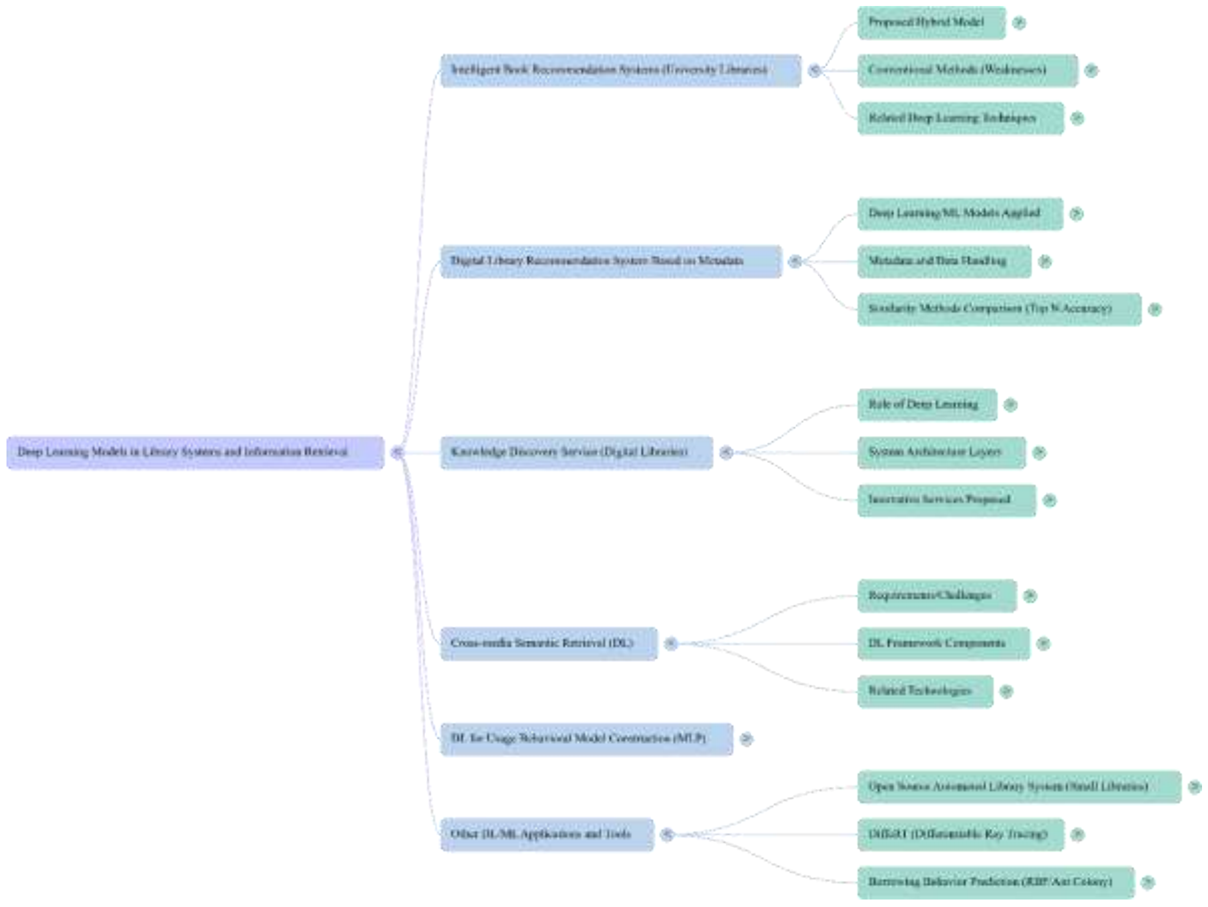

Deep Learning Models in Library Systems and Information Retrieval
Intelligent Book Recommendation Systems (University Libraries)
Proposed Hybrid Model
Conventional Methods (Weaknesses)
Related Deep Learning Techniques
Digital Library Recommendation System Based on Metadata
Deep Learning/ML Models Applied
Metadata and Data Handling
Similarity Methods Comparison (Top N Accuracy)
Knowledge Discovery Service (Digital Libraries)
Role of Deep Learning
System Architecture Layers
Innovative Services Proposed
Cross-media Semantic Retrieval (DL)
Requirements/Challenges
DL Framework Components
Related Technologies
DL for Usage Behavioral Model Construction (MLP)
Other DL/ML Applications and Tools
Open Source Automated Library System (Small Libraries)
DiffiRT (Differentiable Ray Tracing)
Borrowing Behavior Prediction (RBF/Ant Colony)

Figure 1 Overall summary of AI in library

3.Results and Discussions

The boxplot illustrates the distribution of borrow counts across all books. The central box shows that the majority of books were borrowed only a small number of times, indicating low overall circulation. Numerous points above the box represent high-borrow outliers, reflecting a small subset of titles that were borrowed significantly more frequently than the rest.

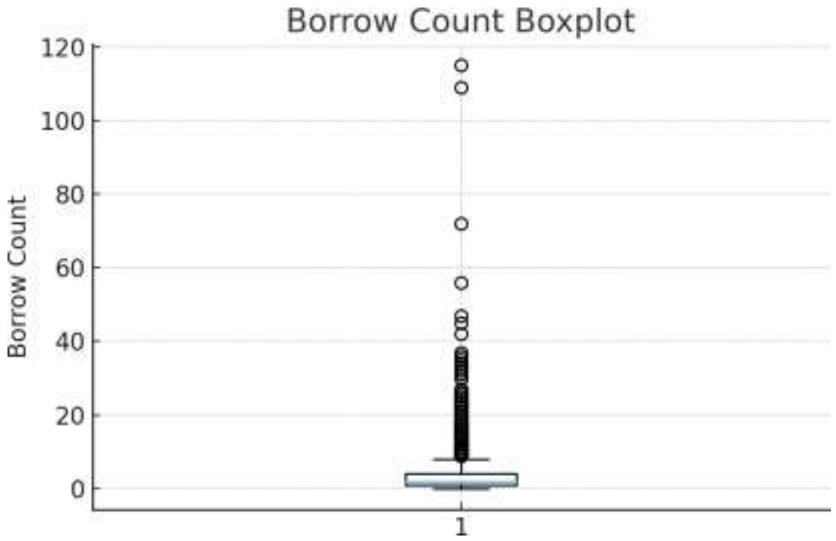


Figure 2 Borrow count

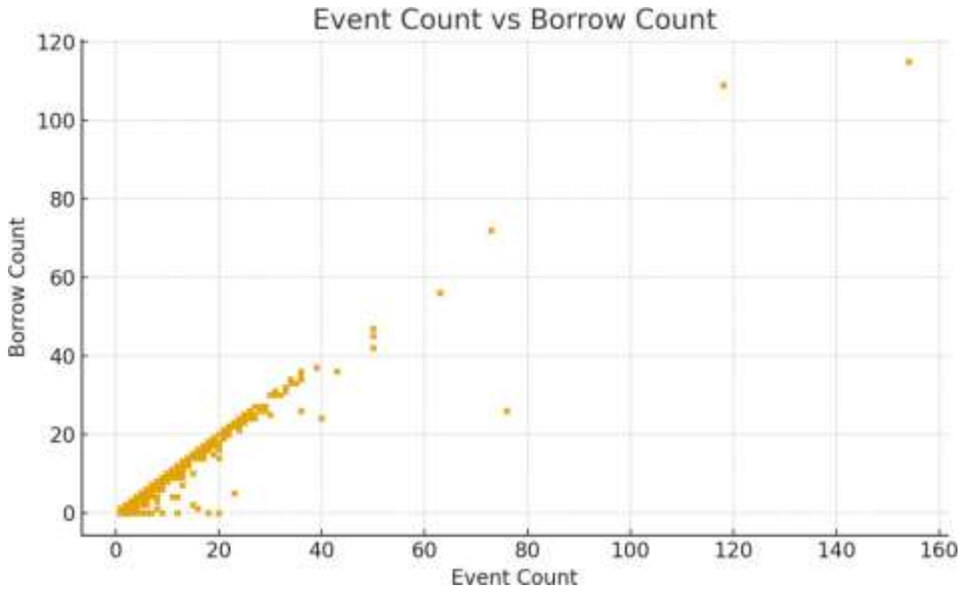


Figure 3 Event count

This scatter plot shows the relationship between event count and borrow count for each book. The upward trend indicates that books with more recorded events generally exhibit higher borrow counts. A cluster of points near the lower values suggests most books experience limited activity, while a few points at higher values reflect titles with exceptionally high engagement.

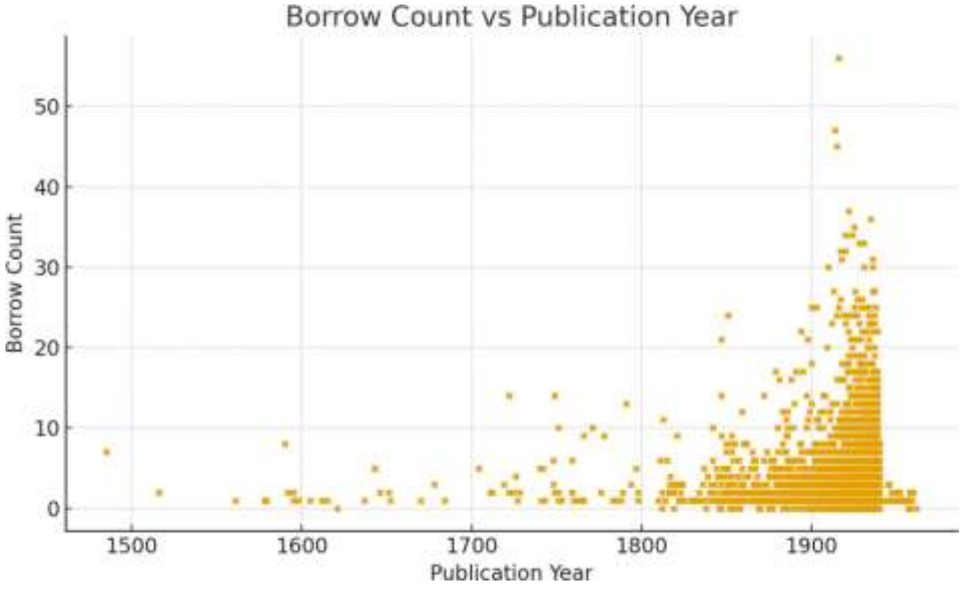


Figure 4 Publication in years

This scatter plot displays the relationship between a book's publication year and its borrow count. Borrow activity is minimal for older publications, with a noticeable increase in both the number of titles and borrowing frequency for books published after the mid-1800s. The concentration of higher borrow counts in more recent years indicates that newer publications tend to circulate more actively within the collection.

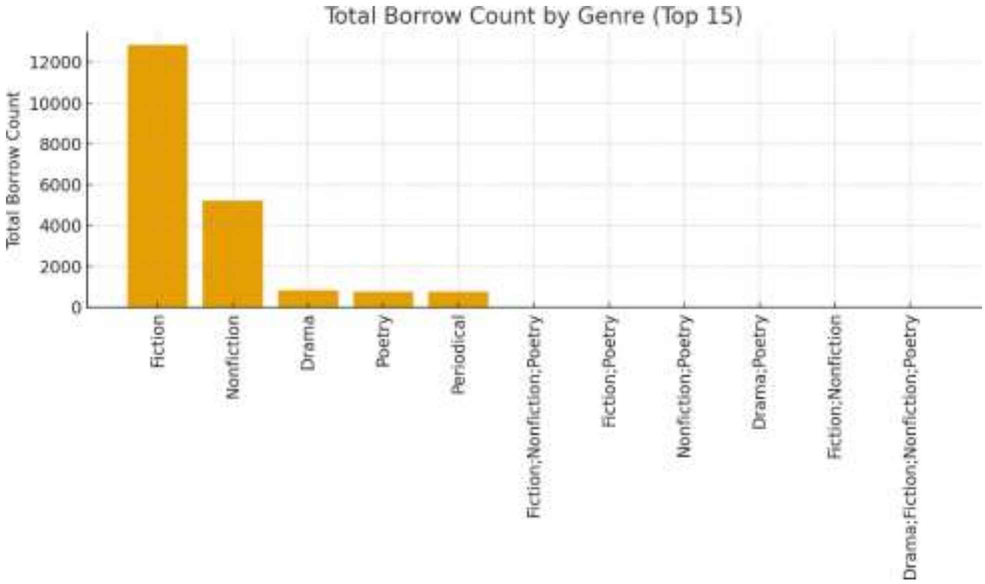


Figure 5 Top lending books

This bar chart presents the total borrow count for the top fifteen genre categories. Fiction overwhelmingly leads in overall borrowing, followed by Nonfiction, while Drama, Poetry, and Periodicals show considerably lower totals. The sharp decline across categories indicates that circulation activity is heavily concentrated within a few dominant genres.

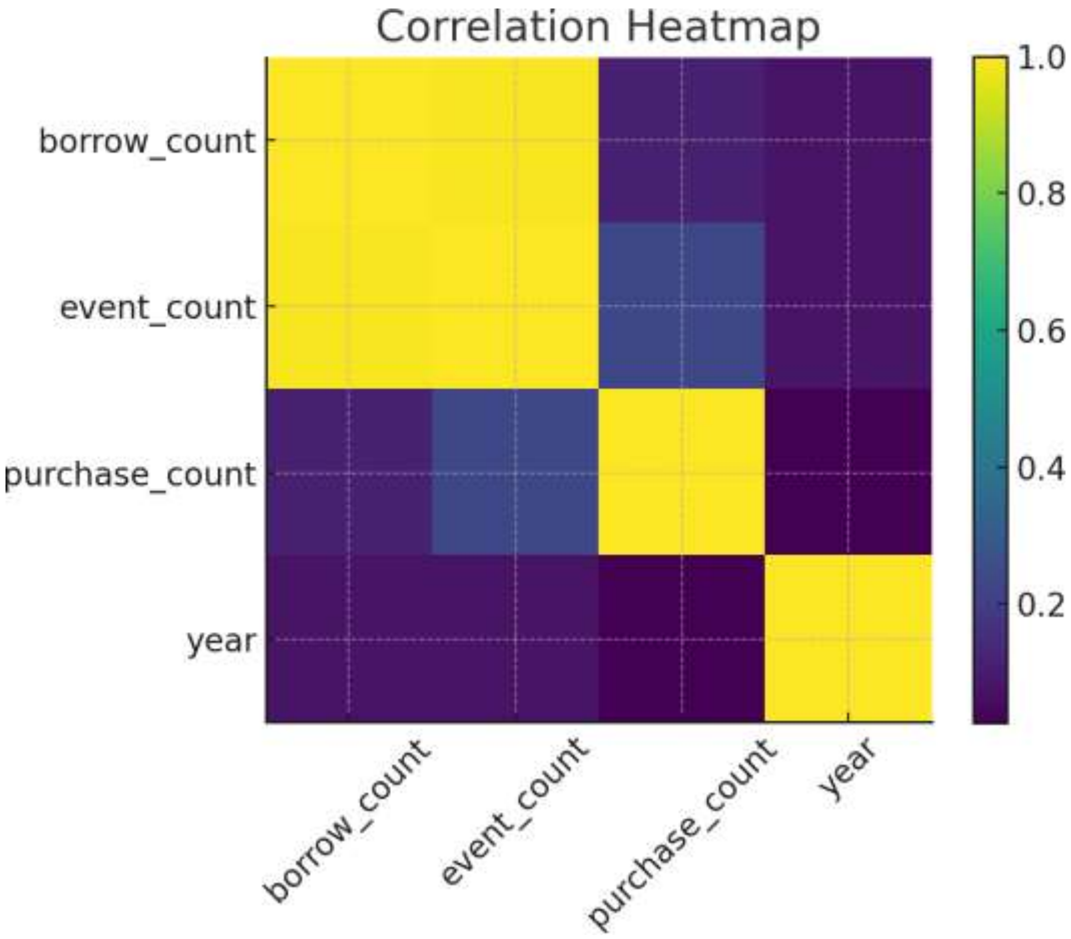
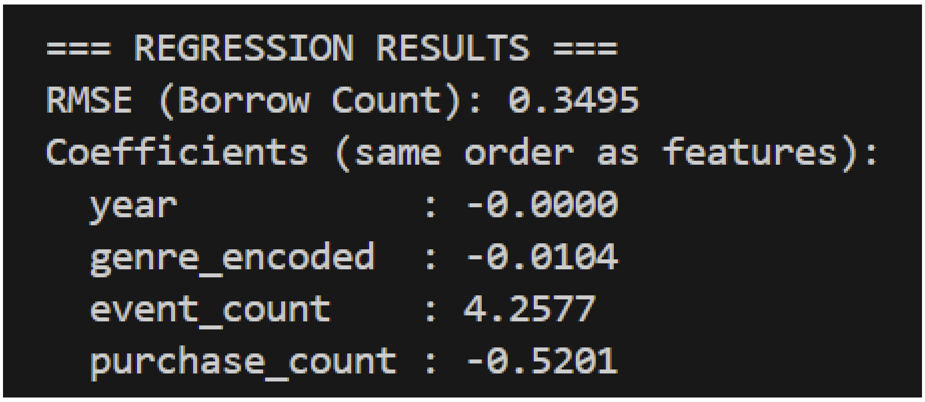


Figure 6 Correlation mapping

The correlation heatmap displays the strength of linear relationships among key numerical variables. Borrow count and event count show a strong positive correlation, indicating that books with more recorded events tend to be borrowed more frequently. Purchase count exhibits only weak correlations with other variables, while publication year has minimal association with all activity measures.

Regression Results:



```
=== REGRESSION RESULTS ===
RMSE (Borrow Count): 0.3495
Coefficients (same order as features):
  year            : -0.0000
  genre_encoded   : -0.0104
  event_count     : 4.2577
  purchase_count  : -0.5201
```

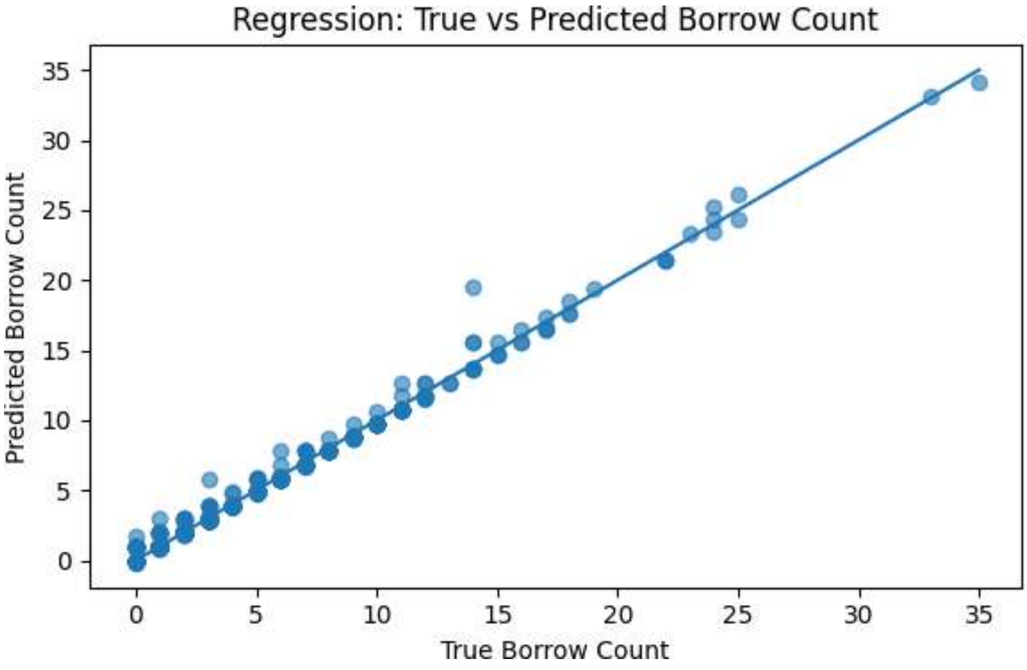
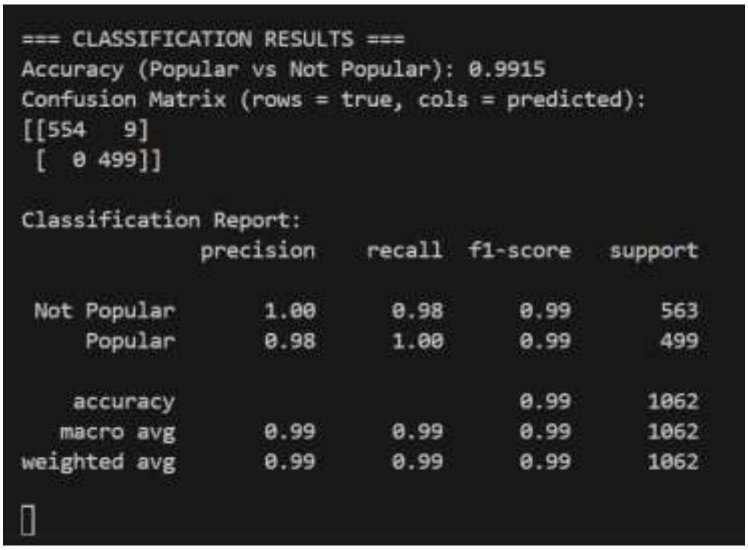


Figure 6 Predicted lending books

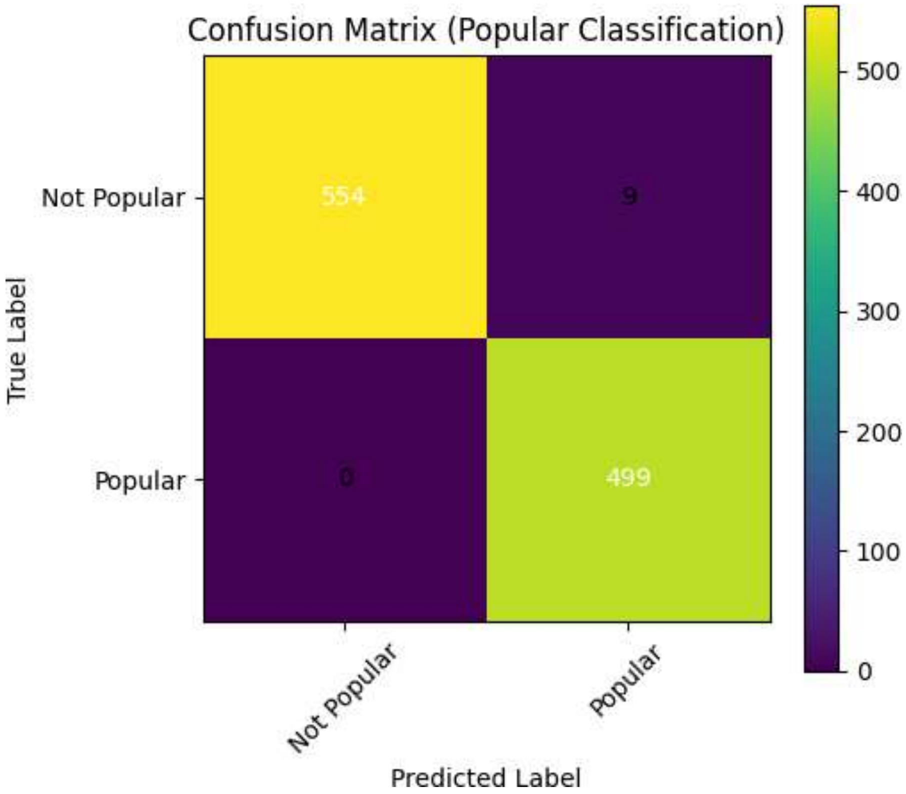
Inference:

The regression results show that event_count is the strongest predictor of borrow_count, meaning books with more recorded events are borrowed far more frequently. The very small or negative coefficients for year, genre, and purchase_count indicate that these factors contribute minimally compared to event activity.

Classification Results:

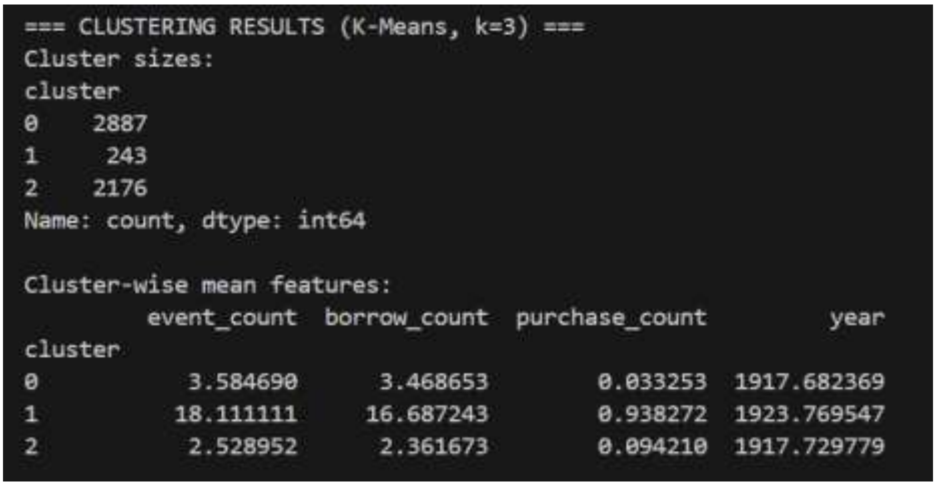


```
=== CLASSIFICATION RESULTS ===
Accuracy (Popular vs Not Popular): 0.9915
Confusion Matrix (rows = true, cols = predicted):
[[554   9]
 [  0 499]]

Classification Report:
              precision    recall  f1-score   support

 Not Popular       1.00      0.98      0.99       563
     Popular       0.98      1.00      0.99       499

    accuracy                           0.99      1062
   macro avg       0.99      0.99      0.99      1062
weighted avg       0.99      0.99      0.99      1062

▯
```

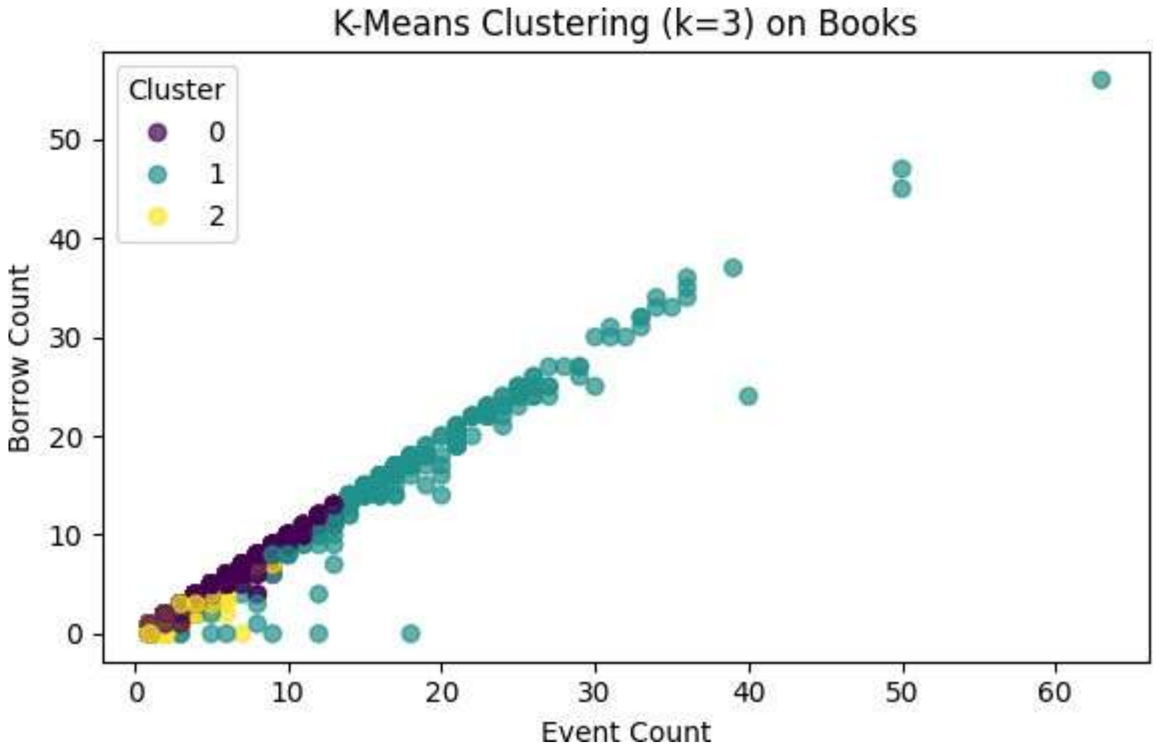


The classification results show that the model can almost perfectly distinguish between popular and non-popular books, with a 99% accuracy rate and very few misclassifications. This indicates that the chosen features—especially event_count—provide a very strong signal for predicting a book's popularity.

K means Clustering Results:

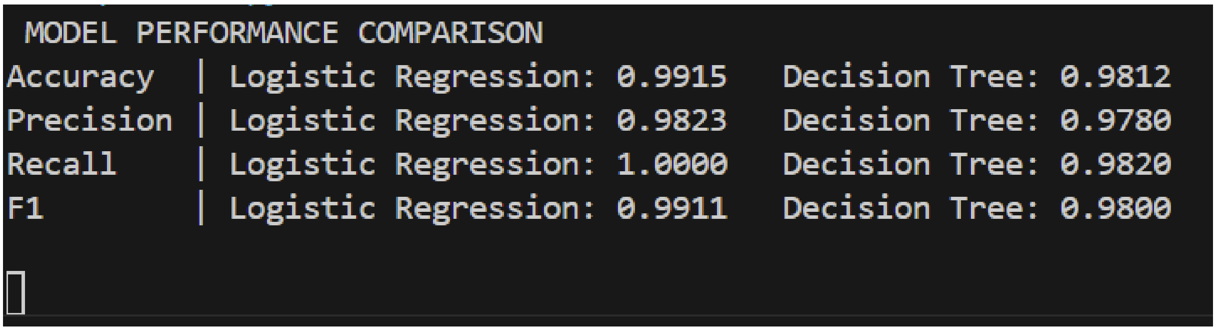


```
=== CLUSTERING RESULTS (K-Means, k=3) ===
Cluster sizes:
cluster
0    2887
1     243
2    2176
Name: count, dtype: int64

Cluster-wise mean features:
         event_count  borrow_count  purchase_count         year
cluster
0           3.584690      3.468653        0.033253  1917.682369
1          18.111111     16.687243        0.938272  1923.769547
2           2.528952      2.361673        0.094210  1917.729779
```

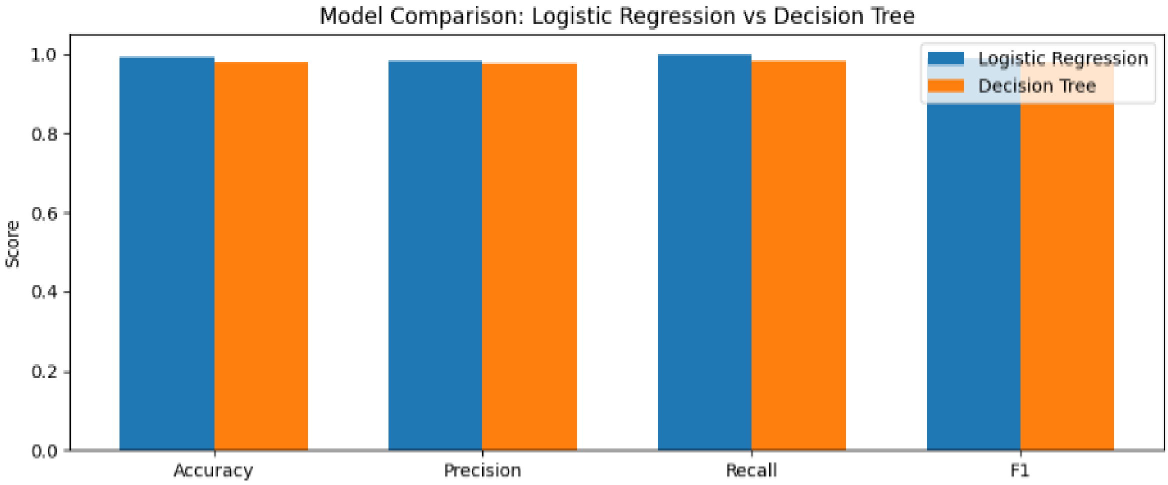


The K-Means model grouped the books into three distinct clusters representing high-usage, medium-usage, and low-usage patterns based on event and borrow activity. Cluster 1 contains the most active books with very high event and borrow counts, while Clusters 0 and 2 represent books with moderate and minimal usage respectively.

Comparison between Logistic regression and Decision Tree Classifier



```
 MODEL PERFORMANCE COMPARISON
Accuracy  | Logistic Regression: 0.9915   Decision Tree: 0.9812
Precision | Logistic Regression: 0.9823   Decision Tree: 0.9780
Recall    | Logistic Regression: 1.0000   Decision Tree: 0.9820
F1        | Logistic Regression: 0.9911   Decision Tree: 0.9800
```



Logistic Regression outperforms the Decision Tree across all major metrics—accuracy, precision, recall, and F1-score. This is because the relationship between the dataset's features (especially event_count) and the target variable (popularity) is highly linear and well- separated, which suits Logistic Regression perfectly, while the Decision Tree introduces unnecessary splits and slightly overfits, resulting in lower recall and F1-score.

4. Conclusion

The study demonstrates that Artificial Intelligence provides powerful tools for analyzing library borrowing data and understanding user behavior. AI-based regression models reveal that **event_count** is the most influential predictor of borrow_count, while other variables, such as publication year and purchase_count, play a comparatively minor role. Classification models confirm that book popularity can be predicted with near-perfect accuracy, with Logistic Regression achieving **99% accuracy**, outperforming Decision Trees due to its robustness in linearly separable datasets. K-Means clustering further distinguishes library materials into clear behavioral groups, offering actionable insights into high-usage and low-usage collections. These findings highlight that AI and machine learning can significantly enhance library management by enabling precise prediction, segmentation, and interpretation of user engagement. By adopting such intelligent analytical frameworks, libraries can improve collection development, tailor user services, and make informed decisions that support evolving reader needs.

References

1. Rui Guo, Yiwei Pang, Yuanxi Xu, Zhenyang Liu, Yanchen Chen, Yajun Guo; Application of artificial intelligence technologies in library services at the top 100 US universities. The Electronic Library 19 August 2025; 43 (4): 619–648

2. Yan, Rumeng & Zhao, Xin & Mazumdar, Suvodeep. (2023). Chatbots in libraries: A systematic literature review. Education for Information. 39. 1-19. 10.3233/EFI-230045.

3. Ian Tai and Souvick Ghosh. 2025. Integrating AI into Library Systems: A Perspective on Applications and Challenges. Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries. Association for Computing Machinery, New York, NY, USA, Article 42, 1–11.

4. Mwantimwa, K., & Msoffe, G. (2025). Application of generative artificial intelligence in library operations and service delivery: A scoping review. Technical Services Quarterly, 42(2), 139–168.

5. Zhang, Qiu, and Min Wan. "An Intelligent Book Recommendation System for University Libraries based on Deep Learning Models." 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), 2025

6. Almaghrabi, Maram, and Girija Chetty. "Deep machine learning digital library recommendation system based on metadata for Arabic and English languages." 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). IEEE, 2020.

7. Eertmans, Jérome, Claude Oestges, and Laurent Jacques. "Demonstrating DiffeRT: An Open-Source Library for Optimizing Radio Networks with Differentiable Ray Tracing." 2025 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN). IEEE, 2025.

8. Puritat, Kitti, and Kannikar Intawong. "Development of an open source automated library system with book recommedation system for small libraries." 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON). IEEE, 2020.

9. Lin, Wei-Hsiang, et al. "Exploration of usage behavioral model construction for university library electronic resources from Deep Learning Multilayer perceptron." 2019 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW). IEEE, 2019.

10. DongNing. "Research on Cross-media Semantic Retrieval in Digital Library Based on Deep Learning." 27-28 October 2023, Istanbul, Turkey

11. Dong, Ning. "Research on Knowledge Discovery Service in Digital Libraries Based on Deep Learning." 2024 13th International Conference on Educational and Information Technology (ICEIT). IEEE, 2024.

12. Xue, Teng, Wang Xianqing, and Liu Tingting. "Research on Personalized Recommendation System of Library Collection Based on Deep Learning." 2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA). IEEE, 2024.

13. Zhang, Xin, and Desheng Tian. "Research on the Prediction Model of University Library Borrowing Behavior Based on Big Data and Machine Learning." 2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA). IEEE, 2024.

14. Balamani, T., et al. "Rural empowerment through smart learning environment." 2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA). IEEE, 2024.