Journal of Applied Engineering and Technological Science
Vol 4(2) 2023: 983-991

FAVORITE BOOK PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS

Dersin Daimari[1], Subhash Mondal[2*], Bihung Brahma[3], Amitava Nag[4]
Dept. of HSS, Central Institute of Technology Kokrajhar[1,3]
Dept. of CSE, Central Institute of Technology Kokrajhar[2,4]
dersindaimari@gmail.com[1], ph22cse1001@cit.ac.in[2*], b.brahma@cit.ac.in[3], amitava.nag@cit.ac.in[4]

Received : 05 April 2023, Revised: 08 May 2023, Accepted : 08 May 2023
**Corresponding Author*

ABSTRACT
Recent years have seen the rapid deployment of Artificial Intelligence (AI) which allows systems to take intelligent decisions. AI breakthroughs could radically change modern libraries' operations. However, introducing AI in modern libraries is a challenging task. This research explores the potential for smart libraries to improve the caliber of user services through the use of machine learning (ML) techniques. The proposed work investigates machine learning methods such as Random Forest (RF) and boosting algorithms, including Light Gradient Boosting Machine (LGBM), Histogram-based gradient boosting (HGB), Extreme gradient boosting (XGB), CatBoost (CB), AdaBoost (AB), and Gradient Boosting (GB) for the task of identifying and classifying Favorite books and compares their performances. Comprehensive experiments performed on the publicly available dataset (Art Garfunkel's Library) show that the proposed model can effectively handle the task of identifying and classifying Favorite books. Experimental results show that LGBM has achieved outstanding performance with an accuracy rate of 94.9367% than Random Forest and other boosting ML algorithms. This empirical research work takes advantage of AI adoption in libraries using machine learning techniques. To the best of our knowledge, we are the first to develop an intelligent application for the modern library to automatically identify and classify Favorite books.
Keywords: *Machine Learning, artificial intelligence (AI), Smart Library, Light Gradient Boosting Machine (LGBM), Boosting Model*

1. Introduction

Libraries have long been one of the most important places to find knowledge. These libraries, formerly only considered information repositories, have a new perspective in the current Information and communication technology (ICT) era. Whether traditional or "smart," libraries aim to share real-time information with their users or members. Information and communication technology (ICT) advancements like artificial intelligence (AI)(Iskandar et al., 2022), machine learning (ML), the Internet of Things (IoT) (Irawan et al., 2022), and fuzzy logic (Perdana et al., 2022) have recently outpaced traditional science and technology in a remarkable way. These new ICTs have the capacity to use analytics and machine learning for trend analysis and prediction, and they can allow modernizations that raise the caliber of services provided by contemporary libraries. This makes it possible for end users to be satisfied. While a computer program designed to mimic human intelligence is the standard definition of artificial intelligence (AI), the truth is much more nuanced. The Google search engine and Chat-GPT are probably the most notable examples of how daily applications of AI are already fully incorporated into society through technology like autonomous vehicles, robotics, intelligent chatbots, and so on.

Machine Learning is a subfield of artificial intelligence that enables machines to make intelligent decisions without human interference through sophisticated mathematical and statistical methods with the help of devices or human-generated data (Garg & Mago, 2021) (Robles Mendo et al., 2021) (Pasa Uysal, 2022). This idea of automating complex tasks has generated high interest in different applications such as smart healthcare, smart city, smart library, etc. Machine learning techniques are mainly classified as supervised and unsupervised learning, distinguished by the presence of data labelling, also known as feedback. When the model is given a dataset and the corresponding label of each data, it learns through supervised learning to recognize the correct label. Figure 1 shows a thematic taxonomy of AI and ML.

The authors (Jayawardena et al., 2021) (Shi et al., 2020) (Yee Chu & L. Wong, 2021) (Tang et al., 2018) (He et al., 2020) uses the ML and deep learning based uses of book title prediction in the library system and the different applications of AI to transform the library system in the smart application.

Machine Learning plays an important role in overcoming the current obstacles faced in today's library system to make it a smart system. In this work, we developed an automated system in order to identify favorite books. The proposed system uses seven state-of-the-art machine learning methods (LGBM, RF, HGB, CB, GB, XGB, AB) by learning data from Art Garfunkel's library dataset. To the best of our knowledge, this is the first work for favorite book identification.

The rest of the paper is organized as follows: In Section 2, we analyze the related works. Section 3 describes the proposed method. Section 4 consists of the experimental results. Finally, in Section 5, we conclude

2. Literature Review

The authors (Hervieux & Wheatley, 2021) discussed the survey of the academic librarian's perception of the implementation of AI in libraries. The survey concluded that the librarians do not fully agree with incorporating AI in an academic library, nor have proper knowledge about modern technology. They also require training to familiarize themselves with the implication of AI in the library system. The authors provide a proposal to implement AI to improve the workflow of the library service. The authors (Yoon , E. Andrews, & L. Ward, 2022) (Harisanty et al., 2022) (Tait & M Pierson, 2022) (Hamad et al., 2023) (M. Cox & Mazumdar, 2022) are discussed about the usages of AI in library management by the public and the leaders in perspective of many countries.

In (Yousuf Ali et al, 2020) focused on the implication of Natural Language Processing (NLP) in the university library and the different AI-based tools used to cope with the modern technology implementation in the library system. Technological skills and knowledge are required to implementation of modern AI tools. In the research library, the uses of AI and its implementation are summarized by many researchers (Andrea Alessandro & Heli, 2022) (Ullah et al., 2022) (Hussain, 2023) (Emmanuel Ajakaye, 2022) (Hassan Azimi et al., 2022).

The authors (Abayomi et al., 2020) investigated a survey of AI-based awareness keeping in the new qualitative and quantitative perception approach about the library management system. Investigate found that they are aware of the implications of AI technology but fear job loss. The academic librarian needs more skills in adopting technologies to provide end-user satisfaction in the library management system.

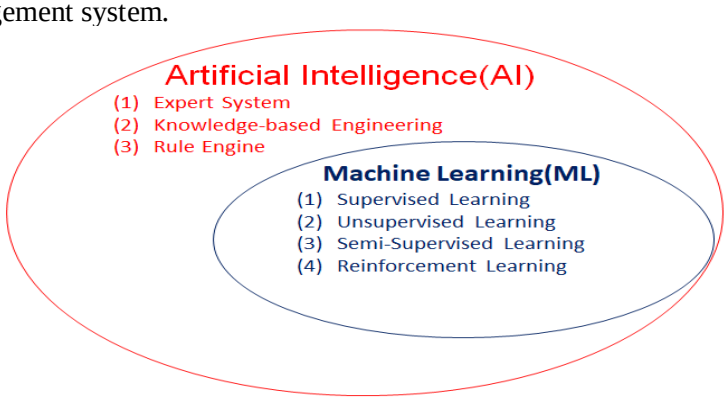

Fig. 1. Thematic Taxonomy of AI and ML

The authors (Folorunsho & Omeiza Momoh, 2020) focused on the use of AI and ML technologies in library operations, encouraging human thought and improving utilitarian library use. In developed countries, many university libraries have begun implementing AI and ML, however, in most developing countries. In (Tsabedze et al., 2022) addressed the issues of integrating AI into the library information system curriculum, which provided the momentum for the research incorporating modern tools in the library system to be smart in the future.

The authors (Daimari et al., 2021) developed a machine learning-based model to predict the possible date of availability of the already issued books by some users. The framework utilized the library data Central Library of CIT Kokrajhar. In (Geetha et al., 2022) (Mupaikwa, 2023) also

discussed the uses of ML in library management systems. In (Jayawardena et al., 2021) and (Bi et al., 2022) presented an RFID-based and IoT-based smart library management system. The system intelligently controls library shelves, suggest books, reserve seats, etc. A very intriguing positioning system solution based on the integration of IoT (BLE beacons, Wi-Fi) and ML (K-Nearest Neighbor (KNN)) is proposed by (Antevski et al., 2016) to help library users by intelligently optimizing space consumption. As a result, the proposed system improves smart space service efficiency. The authors discuss in the literature (A, 2022) the likelihood of the adoption of different approaches to AI in academic libraries. He shows the different applications of AI for knowledge discovery. In (Bai, 2022) elaborate on the need of library users for a specialized service platform for wisdom library disciplines. In this paper, the authors presented a BACO-based multidoor image segmentation bun type for facial features, as well as a BACO-based continuous ant colony optimization (CACO) method. In (Zeng et al., 2022) (Shi et al., 2023) demonstrated the possible use of a mobile search-based CNN technique used in the library system. The work (Kang, 2022) proposes a Deep Learning based personal book recommendation system of a smart library. The system improves the borrowing rate of books and consequently maximizes the utilization of book resources.

In (Zhang, 2021) (Li et al., 2021) (Shi & Zhu, 2022) (Meddeb et al., 2021) (Shu et al., 2021) described the current application of 5G era, mobile-based service and big data approaches and the implication the possible uses of AI and ML in smart library management systems.

In this work, we have built seven different machine learning models such as Light Gradient Boosting Machine (LGBM), Histogram-based gradient boosting (HGB), Extreme Gradient Boosting (XGB), CatBoost (CB), AdaBoost (AB), Gradient Boosting (GB), and Random Forest (RF) in order to identify Favorite books based on the data of Art Garfunkel's Library. After comparing the performances of all seven models, we concluded that LGBM provided the highest accuracy.

3. Research Methods

This section discusses the proposed methodology of studying a person's favorite book in a library.

Dataset Acquisition

The dataset used in this study is taken from the open-source library of Garfunkel's Favorite Book (lostinworlds., 2022). The dataset contains 1321 instances with six attributes. The target outcome, "Favorite" had two classes, either favorite class "1" or not favorite class represented in "0". During the dataset analysis phase, it was observed that the target values had 1185 instances of class 0 and 136 instances of class 1. The description of each feature is represented in Table 1.

Table 1- Feature Details of The Dataset

Feature/Column Name	Data Type	Column type	Feature/Column Description
Date Read	object	Categorical	Tells the reading date with the Month and Year
Author	object	Categorical	Tells the name of the Author of the book
Books	object	Categorical	Tells the title of the Books
Year Published	int64	Numerical	It tells the year the book published
Pages	int64	Numerical	Indicates the total number of pages of a book
Favorite	int64	Binary	Represents in binary either "1" Favorite, or "0", not Favorite

Data Pre-processing

The open-source raw dataset generally represented input data with features usually in structured columns. Most of the time, the raw dataset requires many preprocessing techniques to process the data compatible for the training purpose of any prediction model. To gain a better accurate model and eliminate the overfitting problem of the model, perform data preprocessing techniques like removing unwanted noise and outliers and handling missing values of the dataset. To take the categorical column data types, it is required to convert them into numerical values. Sometimes it is needed to overcome the data imbalance of the target outcome using the oversampling technique. During the data analysis phases, it is observed that all the above

problems are present in the dataset. To overcome the above issues, for all the data mentioned above, preprocessing techniques are applied to the dataset before the final model training and testing for prediction to classify a person's favorite book in a library.

Handle Missing Values

A few impossible or missing numbers were in the "Year Published" feature column. After removing the impossible values, the missing values are substituted with the mean values.

Encoding Technique

The One-Hot Encoding technique was applied to convert the categorical data into an integer without considering the dummy variable trap to handle the string-type categorical features. Also, the label encoder technique was used to convert the object-type data into unique numerical values. Three columns, namely "Author," "Books," and "Date Read," are handled with the encoding technique.

Balancing Imbalanced Data

It is observed in the data analysis process that the output column was highly imbalanced. To balance the entire column with an equal number of classes, the oversampling technique SMOTE was applied to sampling the minority class. After oversampling, the dataset contains a total of 2370 instances. The values pre- and post-balancing are presented in Table 2.

Table 2 - Imbalance Data Balancing

	Before SMOTE	After SMOTE
Label 0 counts from the Favorite column	1185	1185
Label 1 count from the Favorite column	126	1185

Scaling Features to Common Range

Finally, standardizing the dataset attribute values in standard ranges between 0 and 1 to simplify the model training process, the MinMax scalar was applied to the entire dataset of independent variables.

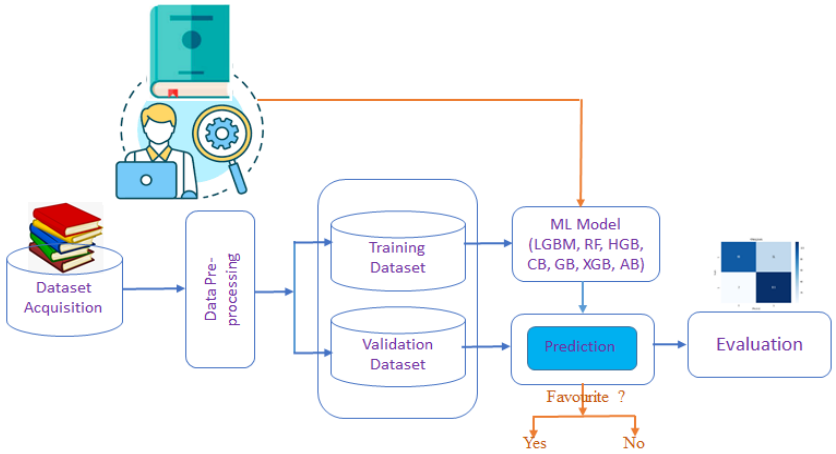

Fig. 2. Detailed Block Diagram of The Proposed Framework

After the dataset has been preprocessed and significant features have been chosen, the all models have been trained using the selected features. The block diagram in Fig. 2 shows all phases of the procedure in their entirety.

4. Results and Discussions

This section reports the performance of all seven models. After applying all data pre-processing techniques discussed in section 3, the dataset was split into the ratio of train and test of 0.90:0.10, respectively. The feature correlation heat map among all features and the target variable is depicted in Fig. 3.

Performance Metrics

The performance has been evaluated using four metrics – Precision (P), Accuracy (A), F1-Score(F), and Recall (R). The confusion matrix is shown in Fig. 4. Given the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), these metrics are evaluated based on the confusion matrix as follows:

Accuracy(A): Accuracy is defined as the proportion of right predictions to the total number of predictions made on a given set of data as given below:

$$Accuracy = \frac{TP + TN}{(TP + FN) + (FP + TN)}$$

Precision(P): It is the ratio of successfully predicted positive data points to the number of data points correctly predicted as positive by the classifier, as expressed by the following formula:

$$Precision = \frac{TP}{(TP + FP)}$$

Recall (R): It's the proportion of accurately anticipated positive data points to the total number of positive data points. It is defined as:

$$Recall = \frac{TP}{(TP + FN)}$$

F1-score (F): The F1-score is used to determine the accuracy of a test set. The harmonic mean of precision and sensitivity is the F1 score.

$$F1 - score = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

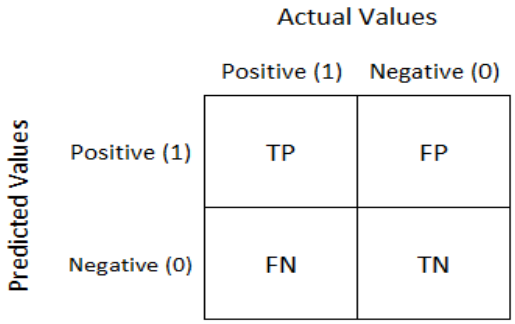

Fig. 3. General Confusion Matrix

Model Training

To build up the prediction model based on machine learning techniques, the selection of the ML model plays a vital role in any classification-based prediction model. The model was deployed by considering all the boosting ensemble techniques like Light Gradient Boosting Machine (LGBM), Histogram-based gradient boosting (HGB), Extreme gradient boosting (XGB), CatBoost (CB), AdaBoost (AB), and Gradient boosting (GB), along with a non-boosting tree-based ensemble technique like Random Forest (RF). The above model was trained with 90% of the total dataset instances; the rest, 10%, was used to test the model performances. Also, the k-fold cross-validation technique was applied to the training dataset to get the mean accuracy of the model. The models were tested with 10-fold cross-validation to produce Mean Accuracy (MA). The model performances were measured by taking the different metrics like Accuracy (A), Precision (P), Recall (R), F1-Score (F), Cohen-Kappa Score (CK), and ROC-AUC score (RA). As the dataset was fewer instances, it is prone to overfitting. To overcome and measure the stability of the model, the standard deviation (SD) was also evaluated for each model. The LGBM

model performed better, with an accuracy of 94.94% and a minimum standard deviation of 0.0171, much less than other boosting models. Also, the RF model performed better with accuracy and a 10-fold mean accuracy of 93.25% and 91.32%. We put the results of the performance measurement metrics in the percentage of all models in Table 3.

The confusion matrix is shown in Figure 5. As can be seen in Figure 5, LGBM, RF, and HGB achieve high classification accuracy. Figure 6 shows the ROC-AUC curve of the investigated models under test conditions. The area under the curve using LGBM is 99%.

Fig. 4. Feature Correlation Heat Map

One major finding of the proposed work is that the LGBM machine learning model can perform better classification than other boosting and RF ML models. As a result, the proposed system using LGBM may be an effective tool to help readers choose a book. Consequently, this strategy can greatly shorten the time required for book selection and encourage smart libraries to adopt automated methods.

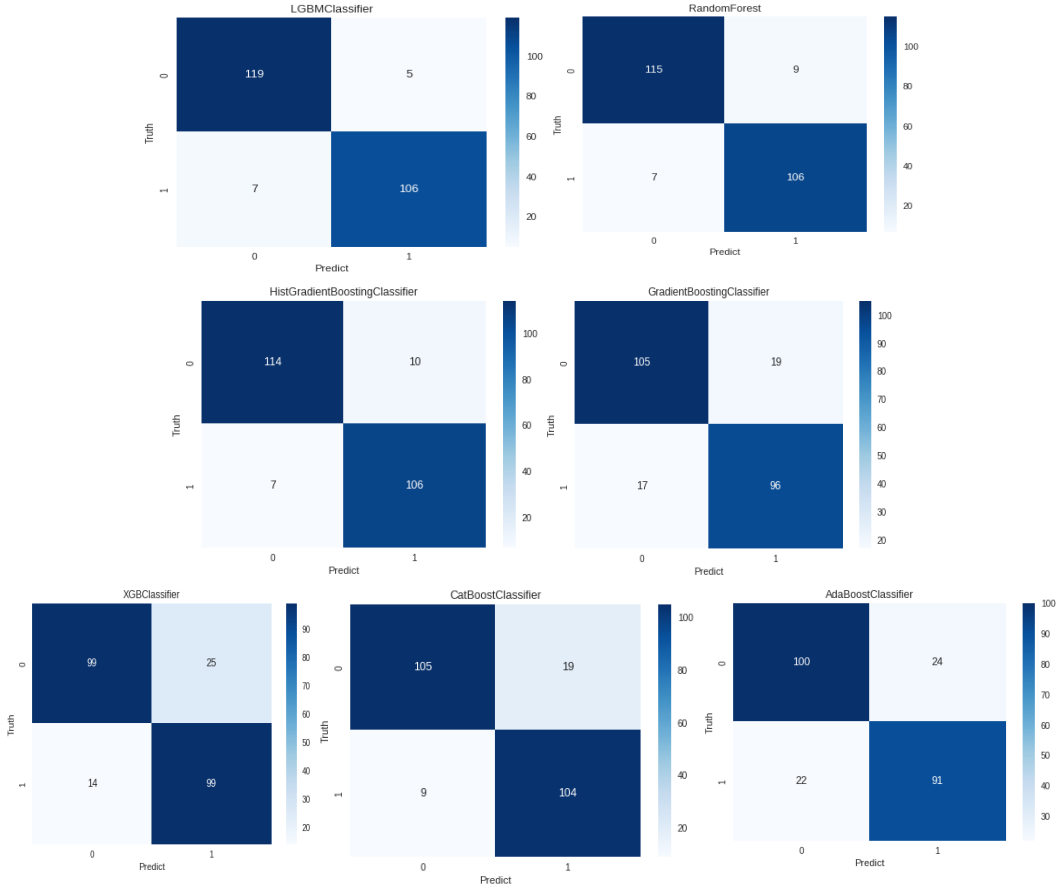

Fig. 5. Confusion Matrix of The Deployed Model

Table 3 - Experimental Results of The Investigated Models

Model	A	MA	SD	RA	P	R	F	CK
LGBM	94.9367	92.7305	0.0172	0.9831	0.9550	0.9381	0.9464	0.8984
RF	93.2489	91.3249	0.0203	0.9734	0.9217	0.9381	0.9298	0.8648
HGB	92.8270	92.2619	0.0225	0.9813	0.9138	0.9381	0.9258	0.8564
CB	88.1857	86.8253	0.0219	0.9409	0.8455	0.9204	0.8814	0.7641
GB	84.8101	87.4352	0.0184	0.9398	0.8348	0.8496	0.8421	0.6958
XGB	83.5443	86.0278	0.0226	0.9296	0.7984	0.8761	0.8354	0.6716
AB	80.5907	78.7157	0.0224	0.8735	0.7913	0.8053	0.7982	0.6113

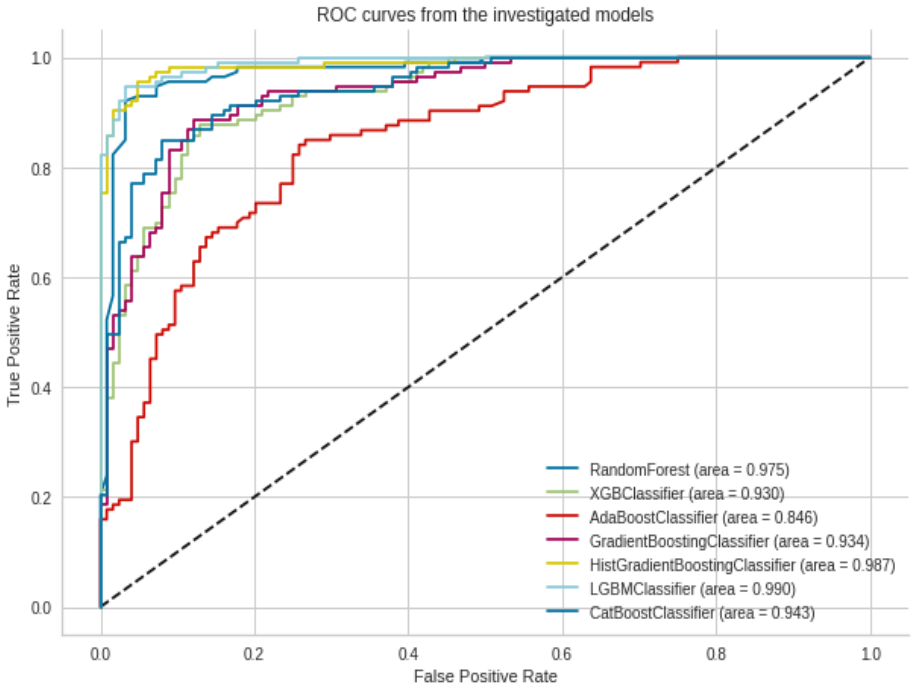

Fig. 6. ROC-AUC curve of the deployed model

5. Conclusion

Seven supervised ML models, LGBM, RF, HGB, CB, GB, XGB, and AB, are used in this research to offer an AI-based application for a smart library. To examine the outcomes of various models side by side, we ran comparison tests using the same data set. The experimental results demonstrated that the LGBM machine learning model could provide superior classification performance compared to the Random Forest (RF) and other boosting ML models, which is the key finding of this research work. As a result, the presented framework in this paper may be a helpful tool to aid readers in the process of choosing a book. Therefore, this strategy can greatly aid in cutting down on time required for book selection and encourage smart libraries to implement machine learning-based automated methods. To satisfy end users, we want to create a mobile/web application in the future for library patrons.

References

A, C. (2022). How artificial intelligence might change academic library work: Applying the competencies literature and the theory of the professions. *Journal of the Association for Information Science and Technology*, *74*(3), 367-380. doi:https://doi.org/10.1002/asi.24635

Abayomi, O., Adenekan, F., & Adeleke, O. (2020). Awareness and Perception of the Artificial Intelligence in the Management of University Libraries in Nigeria. *Journal of Interlibrary Loan, Document Delivery & Electronic Reserve*, *29*(2). doi:https://doi.org/10.1080/1072303X.2021.1918602

Andrea Alessandro, G., & Heli, K. (2022). Understanding Artificial Intelligence in Research Libraries: An Extensive Literature Review. *Te Journal of European Research Libraries, 32*(1), 1-36. doi:https://doi.org/10.53377/lq.10934

Antevski, K., E. C. Redondi, A., & Pitic, R. (2016). A hybrid BLE and Wi-Fi localization system for the creation of study groups in smart libraries. *9th IFIP Wireless and Mobile Networking Conference (WMNC).* Colmar, France. doi:10.1109/WMNC.2016.7543928

Bai, Y. (2022). Construction of a Smart Library Subject Precise Service Platform Based on User Needs. *Mathematical Problems in Engineering*, 1-8. doi:DOI: 10.1155/2022/5675291

Bi , S., Wang, C., Zhang , J., & Huang , W. (2022). A Survey on Artificial Intelligence Aided Internet-of-Things Technologies in Emerging Smart Libraries. *sensors, MDPI, 22*(8), 2991. doi:https://doi.org/10.3390/s22082991

Daimari, D., Narzary, M., Mazumdar, N., & Nag, A. (2021). A Machine Learning Based Book Availability Prediction Model for Library Management System. *Library Philosophy and Practice (e-journal)*.

Emmanuel Ajakaye, J. (2022). Applications of Artificial Intelligence (AI) in Libraries. *Handbook of Research on Emerging Trends and Technologies in Librarianship*, 1-18. doi:10.4018/978-1-7998-9094-2.ch006

Folorunsho, A., & Omeiza Momoh, E. (2020). Application of Artificial Intelligence and Robotics in Libraries : A Review of Literature. *ILIS Journal of Librarianship and Informatics, 3*(2), 93 – 98.

Garg, A., & Mago, V. (2021). Role of machine learning in medical research: A survey. *Computer Science Review, 40*. doi:https://doi.org/10.1016/j.cosrev.2021.100370

Geetha, K., Srinivasa Rao, G., Kaur, C., & Kumar, K. (2022). Machine learning based library management system. *6th International Conference on Electronics, Communication and Aerospace Technology*, (pp. 1031-1034). Coimbatore, India. doi:10.1109/ICECA55336.2022.10009423

Hamad, F., Al-Fadel, M., & Khafaga Shehata, A. (2023). The level of digital competencies for the provision of smart information service at academic libraries in Jordan. *Global Knowledge, Memory and Communication*. doi:https://doi.org/10.1108/GKMC-06-2022-0131

Harisanty , D., Variant Anna, N., Eranti Putr, T., & Firdaus, A. (2022). Leaders, practitioners and scientists' awareness of artificial intelligence in libraries: a pilot study. *Library Hi Tech*. doi:https://doi.org/10.1108/LHT-10-2021-0356

Hassan Azimi, M., Mohammadi, Z., & Rafieinasab, F. (2022). A Survey of Academic Librarians' Perceptions of Artificial Intelligence Technology: A Case Study (Librarians of Shahid Chamran University of Ahvaz and Jundishapur University of Medical Sciences). *Library and Information Sciences, 24*(4), 154-177. doi:10.30481/LIS.2021.286969.1831

He, C., Li, S., So, J., & Zeng, X. (2020). FedML: A Research Library and Benchmark for Federated Machine Learning. *Machine Learning*. doi:https://doi.org/10.48550/arXiv.2007.13518

Hervieux, S., & Wheatley, A. (2021). Perceptions of artificial intelligence: A survey of academic librarians in Canada and the United States. *The Journal of Academic Librarianship, 47*(1). doi:https://doi.org/10.1016/j.acalib.2020.102270

Hussain, A. (2023). Use of artificial intelligence in the library services: prospects and challenges. *Library Hi Tech News, 40*(2), 15-17. doi:https://doi.org/10.1108/LHTN-11-2022-0125

Irawan, Y., Sabna, E., Fauzan Azim, A., & Wahyuni, R. (2022). Automatic Chili Plant Watering Based On Internet Of Things (IoT). *Journal of Applied Engineering and Technological Science (JAETS), 3*(2), 77–83. doi: https://doi.org/10.37385/jaets.v3i2.532

Iskandar, . D. M., Yel, M. B., & Maheswara, E. (2022). Sign Language Detection System Using Adaptive Neuro Fuzzy Inference System (ANFIS) Method. *Journal of Applied Engineering and Technological Science (JAETS), 4*(1), 158–167. https://doi.org/10.37385/jaets.v4i1.967

Jayawardena, C., Reyal, S., Kekirideniya, K., & Wijayawardhana, G. (2021). Artificial Intelligence Based Smart Library Management System. *6th IEEE International Conference*

on Recent Advances and Innovations in Engineering (ICRAIE) (pp. 1-6). Kedah, Malaysia: IEEE. doi:10.1109/ICRAIE52900.2021.9703998

Li, J., Liu, Y., & Wang, L. (2021). Design and Development of Promotion APP of University Smart Library Service Platform Based on Network Teaching. *Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (pp. 1344-1347). Palladam, India: IEEE. doi:10.1109/I-SMAC52330.2021.9640779

Tang, J., Wang, Z., & Lei, L. (2018). Book title recognition for smart library with deep learning. *Mobile Multimedia/Image Processing, Security, and Applications*. doi: https://doi.org/10.1117/12.2312245

Tsabedze, V., Mathabela, N., & Ademola, S. (2022). "A Framework for Integrating Artificial Intelligence into Library and Information Science Curricula. *Innovative Technologies for Enhancing Knowledge Access in Academic Libraries*, 1-14. doi:DOI: 10.4018/978-1-6684-3364-5.ch014