Volume 12. No. 3
December 2024

ISSN-2347-2227
E-ISSN-2347-6141
Subscribers copy
Not for sale

i-manager's

Journal on Computer Science

Disseminating new ideas in Information and Computation

i-manager® Publications

ISSN: 2347-2227
Volume 12. No. 3
December 2024

i-manager's

Journal on Computer Science

About the Journal

i-manager's Journal on Computer Science deals with all aspects of computer science and contributes theoretical results and offers a compilation of high quality articles to encompass a wide spectrum of advancements in the actively developed domain. i-manager's Journal on Computer Science covers a great deal of what has been done in the field recently and intends to bring together the most recent advances and applications in all branches of the academic computer science community with new knowledge and technology for the benefit of students, professionals and industrial practitioners.

i-manager's Journal on Computer Science is presently in its 12th Year. The first issue was launched in 2013.

i-manager's Journal on Computer Science is published by i-manager Publications, one of India's leading Academic Journal Publisher, publishing 49 Academic Journals in diverse fields of Engineering, Education, Management and Science.

Why Publish with us

i-manager Publications currently publishes academic Journals in Education, Engineering, Scientific and Management streams. All of i-manager's Journals are supported by highly qualified Editorial Board members who help in presenting high quality content issue after issue. We follow stringent Double Blind Peer Review process to maintain the high quality of our Journals. Our Journals target both Indian as well as International researchers and serve as a medium for knowledge transfer between the developed and developing countries. The Journals have a good mix of International and Indian academic contributions, with the peer-review committee set up with International Educators.

Submission Procedure

First time authors are requested to register themselves on the registration page in the website. Once registered, authors can submit and know the status of their papers.

Researchers and practitioners are expected to submit on or before the stipulated deadline. Authors of accepted proposals will be notified before the stipulated deadline about the status of their proposals. All submitted articles in full text are expected to be submitted before the stipulated deadline.

Review Procedure

All submissions will undergo an abstract review and a double blind review on the full papers.

The abstracts would be reviewed initially and the acceptance and rejection of the abstracts would be notified to the corresponding authors. Once the authors submit the full papers in accordance to the suggestions in the abstract review report, the papers would be forwarded for double-blind review by the review panel. The final selection of the papers would be based on the report of the review panel members, and the approval of the Editor-in-Chief.

Format for Citing Papers

Author surname, initials (s.) (2024). Title of paper. i-manager's Journal on Computer Science, 12(3), xx-xx.

Copyright

Contact e-mails

editor@imanagerjournals.com
submissions@imanagerpublications.com

ISSN: 2347-2227
Volume 12. No. 3
December 2024

i-manager's
Journal on Computer Science

ISSN: 2347-2227
Volume 12. No. 3
December 2024

i-manager's
Journal on Computer Science

OUR OFFICES

Registered Office

3/343,Hill view,
Town Railway Nager,
Nagercoil, Kanyakumari District - 629001
Ph : 91-4652- 277675
E-mail : info@imanagerpublications.com

Editorial Office

13-B, Popular Building,
Mead Street, College Road,
Nagercoil, Kanyakumari District - 629001
Ph : (91-4652) 231675, 232675, 276675
E-mail : editor@imanagerjournals.com

Join with us

https://www.facebook.com/imanJCOM

https://twitter.com/imanagerpub

Abstracting / Indexing

CONTENTS

RESEARCH PAPERS

EDITORIAL

i-manager's Journal on Computer Science (JCOM), December 2024, Volume 12 Issue 3, features five peer-reviewed papers covering various subjects within the realm of Computer Science. The Journal is dedicated to publishing research articles on information and computation, as well as practical techniques for their implementation and application in computer systems.

Umar Danjuma et al. address the issue of scalability in blockchains, providing a comparative analysis of various blockchain metrics using real-time data. This study employs BlockSim to evaluate performance and examines effective techniques to mitigate scalability limitations by comparing simulation results with real-world scenarios. It proposes an efficient algorithm that enhances the scalability of the Bitcoin network through optimized transaction deferment and improves existing Bitcoin protocols by incorporating inventory messaging and transaction adjournment strategies.

Enesi Femi et al. develop an online database of essay questions designed for quick and convenient access by tutors. This system is built using an iterative software engineering approach. A variety of data sources are considered in the creation of the questions. Techniques such as the Fisher-Yates Shuffle algorithm and generative artificial intelligence are explored for curating these data sources. This study specifically focuses on utilizing the upper-level curriculum of the Computer Science Department at the Federal University of Technology, Minna, Nigeria.

Bryan et al. propose a novel algorithm for generating robust passwords, offering an alternative to traditional random password generators. Among the various authentication systems designed to protect data, password-based authentication remains one of the most widely used methods. This study leverages user-provided information to create memorable passwords. The system is tested using diverse synthetic input data, and the strength of the generated passwords is evaluated with four popular online password checkers.

Sarita and Vaibhav propose a novel predictive pipeline that leverages an SVM model optimized using GA to enhance the accuracy and reliability of B-cell epitope predictions. In the field of immunoinformatics, precise B-cell epitope prediction is crucial for vaccine development and therapeutic interventions. This study incorporates a data preprocessing framework, including labeling, normalization, and dataset splitting. The performance of the optimized SVM model is rigorously compared with traditional methods such as Random Forest and K-Nearest Neighbors.

Shruti et al. aim to bridge the communication gap by developing a sensor-embedded glove that translates user gestures into Morse code, a universally recognized system. The Morse code is then converted into auditory and visual feedback, enabling real-time and seamless communication. This device integrates multiple hardware and software components to create a functional and cohesive prototype. The system is designed with a focus on robustness and efficiency to ensure reliable performance. Key hardware elements are responsible for delivering auditory and visual feedback.

We extend our profound thanks to the authors for their contribution towards this issue and we are grateful to the reviewers for spending their quality time in reviewing these papers. Our special thanks to the Editor-in-Chief Dr. Kamal Kumar Mehta for his constant support and efforts in further enhancing the quality of the Journal.

Hope this issue imparts an enlightening reading experience! Enjoy Reading!

Warm regards,

Sahaya Nijuba S.
Associate Editor
i-manager Publications

DATA LAKE SYSTEM FOR ESSAY-BASED QUESTIONS: A SCENARIO FOR THE COMPUTER SCIENCE CURRICULUM

By

ENESI FEMI AMINU * OLUWASEUN ADENIYI OJERINDE ** AYOBAMI EKUNDAYO ***

EZRA ADAMU **** ABISOYE OPEYEMI ADERIIKE *****

HUSSAINI ABUBAKAR ZUBAIRU ****** LAWAL OLAMILEKAN LAWAL *******

*, ***, ****, ******* *Department of Computer Science, School of Information and Communication Technology, Federal University of Technology, Minna, Nigeria.*
** *Department of Computer Science, School of Information and Computer Technology, Federal University of Technology, Minna, Nigeria.*
***** *Department of Computer Science, Federal University of Technology, Minna, Nigeria.*
****** *Department of Information Technology, Federal University of Technology, Minna, Nigeria.*

https://doi.org/10.26634/jcom.12.3.21132

Date Received: 18/08/2024 *Date Revised: 05/10/2024* *Date Accepted: 27/11/2024*

ABSTRACT

Essay questions are a common form of evaluation that instructors and tutors use at all learning levels to gauge students' comprehension skills in relation to the material covered in the course or curriculum. Instructors encounter challenges when constructing questions that are impartial, often framing them under time constraints, which may cause students to perform poorly. This study introduces EQDL, an online database of essay questions that tutors may quickly and conveniently access. An iterative software engineering concept is utilized to accomplish this goal. A variety of data sources, including the ACM or IEEE computer science curriculum, were considered when creating the questions. The Fisher-Yates Shuffle algorithm and generative artificial intelligence were two of the approaches used for gathering data sources. This study uses the Computer Science Department of the Federal University of Technology, Minna, Nigeria's upper levels of the curriculum, particularly the final year, as a scenario. Web-based technologies were employed to design the front end, while the back end was designed using MySQL and other supporting libraries. The proposed system reduces the burden of crafting entirely new questions since data lake questions are easily accessible. Researchers might consider the curriculum for other levels to expand the scope and volume of data in the lake. Furthermore, more advanced technologies can be used to create a mobile-based system.

Keywords: Data Lake, Computer Science Curriculum, Essay Questions, Fisher-Yates Shuffle, Generative Artificial Intelligence.

INTRODUCTION

Organizations recognize the critical role data plays in cutting-edge business intelligence platforms, enabling defensible decision-making and maintaining a competitive edge (Nambiar & Mundra, 2022; Errami et al., 2023). Data lakes are becoming increasingly prominent in big data management and analytics. Unlike traditional "schema-on-write" systems such as data warehouses, data lakes serve as repositories that store raw data in its original formats while providing a uniform access interface (Hai et al., 2023).

To address data access issues, data lakes (DLs) have been proposed as comprehensive repositories that

absorb raw data from heterogeneous sources in their original formats. Unlike data warehouses, which impose a unified schema, DLs require extracting information from raw data sources as a core functionality. Since DLs are a relatively new concept, their designs and functionalities remain topics of ongoing discussion (Hai et al., 2018).

Data lakes allow the storage of enormous quantities of structured, semi-structured, and unstructured data without limitations on account size or file size. They provide a space to store any type of content in its native format. By integrating natively and supporting large data volumes, data lakes enhance analytical performance (Kuppusamy & Suresh Joseph, 2022).

In the field of educational technology, there is growing interest in leveraging technology to foster creativity in the classroom, as numerous studies have shown its favorable impact on learning outcomes (Shabani et al., 2023). Teachers prefer essay questions for assessment because they require students to create answers rather than select them. Essay questions are instrumental in evaluating students' abilities to think critically, analyze, synthesize, and evaluate. They also assess students' understanding of the topic, writing skills, and intellectual engagement. Examinations play a vital role in fostering deeper learning and knowledge acquisition, making it essential to have well-structured examination papers and formats (Bhirangi & Bhoir, 2016; Kumar et al., 2024).

Information retrieval (IR) includes a specialized domain known as question answering (QA). QA systems aim to respond to natural language inquiries with pertinent information. These systems comprise three primary modules: question classification, information retrieval, and answer extraction, along with supporting components (Allam & Haggag, 2012; Zaib et al., 2022).

Developing a variety of questions and themes from diverse perspectives is crucial for implementing a Question Data Lake System. This process involves structuring queries and ensuring secure storage for future use. By leveraging a vast pool of significant questions, a question data lake minimizes cheating, as students are unlikely to encounter identical questions in subsequent rounds. Furthermore, questions are presented in random order, enhancing fairness. The question data lake also supports the educational needs of students. For text generation, various transformer architectures can be utilized (Raffel et al., 2020). The rapid pace of technological development underscores the substantial role of information technology in learning today (Haryanto, 2023).

In response to these developments, this study aims to develop a system that stores raw questions in diverse formats within a repository called a Data Lake. The data lake framework is utilized due to its ability to accommodate heterogeneous data, including structured, semi-structured, and unstructured formats. The question bank contains questions aligned with the ACM or IEEE computer science curriculum, covering courses such as advanced database systems, system simulation and modeling, advanced computer networks, artificial intelligence, computer installation and maintenance, and data mining.

The designed web-based data lake offers refined or processed data as needed, enabling the extraction of significant value from unlimited data types, flexible querying, efficient data management, and the application of diverse tools for insight generation. The proposed system is adaptable and scalable, with the potential to be tailored for other academic disciplines beyond information technology, such as social sciences and the arts, for high school subjects.

1. Related Works

Essien (2023) emphasized salient characteristics, advantages, and effects on educational institutions. To effectively create excellent question papers suited to certain subjects, courses, and assessment levels, QPGS integrates sophisticated algorithms, database management, and user-friendly interfaces. The Fisher-Yates shuffle method, referred to as the Durstenfeld shuffle or the Knuth shuffle, was utilized to efficiently randomize and mix the questions in the question bank. For the system's design, a mix of the Unified Modeling Language (UML) and the Structured System Analysis Methodology

(SSADM) with Object-Oriented Analysis and Design (OOAD) was used. The pitfall of this study is that the algorithm ignores the questions' content and degree of difficulty, which could result in an uneven distribution of the questions presented to users.

The main method used in automated tests was natural language processing with an emphasis on students' written responses. In contrast, mathematical problems frequently require solutions that include visuals like charts, number lines, and geometric shapes. Students can upload their handwritten responses for grading in a number of computer-based learning programs. Li et al. (2024) expanded BLIP, a multi-modal visual reasoning model, in order to construct a multi-task model to simultaneously output scores and comments utilizing students' multi-modal artifacts (texts and images) as inputs. The study refined and assessed this method using a dataset pertaining to open-ended questions and student responses, benchmarked against three baselines. In the study, feedback performance is improved when graphics are included with text inputs. The study's drawback is that student image quality fluctuates, which restricts the model's capacity for recognition.

Clariana and Solnosky (2023) investigated the conceptual quality of summary essays by comparing two conditions: essay prompts with or without a list of 13 broad concepts. A term frequency analysis of the essays indicated that the most network-central concepts appeared with greater frequency, while the other term frequencies were remarkably similar for both the "terms" and "no terms" groups.

This finding suggests a comparable underlying conceptual mental model for the lesson content in both conditions. The quality of the essay networks generated by both AI systems was equivalent to that of the students' essay networks for the broad concepts and the no-concept treatments. A key limitation of the study is the lack of true random assignment to groups; thus, the findings should be interpreted with caution and not overgeneralized. Additionally, there is no well-established method for generating the optimal expert network of lesson content, which constrains both the utility of the expert network as structural feedback and potentially confounds the post-test measures of conceptual structure that rely on this expert network as a referent.

Jiang et al. (2023) enhanced the capacity of Automated Essay Scoring (AES) models for domain generalization in scenarios where target prompt data is unavailable for training. The paper proposed using a prompt-aware neural AES model to extract a comprehensive representation for essay scoring that incorporates both prompt-specific and prompt-invariant elements. The study introduced a novel disentangled representation learning framework to improve the generalization of representations. In the approach, prompt-specific and prompt-invariant information are separated within the representation through the use of a counterfactual self-training technique and a contrastive norm-angular alignment strategy. However, a significant limitation of the study is the reliance on various heuristic assumptions and designs in data augmentation and counterfactual data building. It remains unclear whether these designs and assumptions are applicable to other datasets and languages.

Based on the principles of machine learning, Lyu (2022) developed a writing assistance scoring system for English as a Foreign Language (EFL) learners to reduce the workload associated with paper grading and improve the process's fairness and accuracy. The data was processed using the pandas library, with data preprocessing carried out through the clean method.

The model was tested, cross-validation was implemented, the data was pre-divided, and the programming of the problem scoring system was further optimized. The autonomous scoring system was constructed using an English teaching recognition module, a feature extraction module, and a scoring module. The programming problem table structure was designed, the English writing auxiliary evaluation program was developed, and the writing assistance scoring system was completed. A significant limitation of this study is that feedback on grammar and vocabulary was more comprehensive than feedback on text organization, idea development, and reasoning. Additionally, most of the

comments lacked specificity and were merely suggestive.

The conventional method of creating tests involves manually crafting question papers, a process that is both time-consuming and labor-intensive. To address this, Thigale et al. (2021) proposed the Automatic Question Paper Generator System, which stores multiple question paper sets and diagrams in a database, allowing for easy retrieval and integration into documents. The system employed a keyword-based shuffling algorithm with randomization to create question papers quickly, securely, and without duplication or repetition. However, a limitation of the study is the challenge of balancing question difficulty levels to align with exam criteria, potentially leading to uneven difficulty among question papers.

In the context of artificial training data for question answering (QA), automatic question generation (QG) has shown promise. Sultan et al. (2020) found that diversity-promoting QG outperforms likelihood maximization techniques like beam search in QA training. The study also demonstrated an inverse relationship between diversity and traditional QG evaluation metrics such as BLEU, ROUGE, and METEOR, proposing an intrinsic measure of QG quality that incorporates diversity and correlates strongly with extrinsic QA evaluations. The main limitation is the difficulty in balancing the trade-offs between question diversity and quality.

Wang et al. (2017) aimed to develop a test question bank system for managing test questions, exam papers, and student examinations. Their systematic approach resulted in an enhanced database system for efficient question bank management. However, the study does not address strategies for maintaining fast database query and retrieval response times, which is a significant shortcoming.

Singhal et al. (2016) proposed a difficulty model for generating questions across disciplines based on user-specified difficulty levels. The study utilized lexicographical ordering and an algorithm to process user-defined factor ordering. A "scenario advice" function was also introduced, enabling users to adjust scenarios during operation. While effective, the model struggles with maintaining consistency in difficulty assessments.

Assessments determine an individual's understanding of a subject, but factors like limited question diversity, overly easy questions, subjective evaluations, and lengthy correction times hinder their effectiveness. Risnasari et al. (2021) developed a computer-based testing (CBT) application featuring essays and multiple-choice questions. The CBT incorporates automatic assessment, question classification using Bloom's Taxonomy, and question randomization through the Fisher-Yates Shuffle algorithm. Essay grading is automated using the Smith-Waterman algorithm, which involves preprocessing, data comparison, and converting similarity proportions into scores. However, combining the Fisher-Yates Shuffle and Smith-Waterman algorithms introduces complexity, posing challenges for implementation and optimization, especially for those with limited technical expertise.

Abdullah (2023) designed a web-based computer-based testing system to simplify grading dual-choice answers. The system uses the Fisher-Yates Shuffle algorithm to reduce fraud by randomizing questions for each student. System development followed the Waterfall model, encompassing phases of analysis, design, coding, and testing. A limitation of this study is the algorithm's inability to generate new questions or accommodate diverse question types, which are crucial for comprehensive assessments.

Zhang and Takuma (2015) addressed the high costs of workbooks and the limited availability of question collections by developing a Kanji learning system. The system generates customized question sentences for simulated Kanken exams based on the user's selected grade and question count. It evaluated user responses and provided feedback on learning outcomes. However, a significant limitation is the system's difficulty in disambiguating the multiple meanings and pronunciations of Kanji characters, potentially leading to inaccurate questions.

2. Methodology: An Iterative Software Engineering Principle

An iterative software development methodology is employed to develop the web-based question lake. This methodology is chosen due to its ability to revisit and revise any previous stage of development in response to defects, errors, or omissions.

A robust technique utilized in this study to produce a random permutation of a finite sequence is the Fisher–Yates Shuffle algorithm, also known as the Knuth Shuffle. This algorithm is widely used in scenarios where efficient and unbiased randomization is critical. In this study, the implementation of the Fisher-Yates Shuffle algorithm in an essay-based question data lake system enhances the heterogeneity and stochastic nature of the questions presented to users, ensuring a more equitable evaluation process.

The Fisher–Yates Shuffle algorithm is applied. A large collection of essay-based questions is assembled and stored in the data lake along with metadata, such as topic, difficulty level, and tags, to facilitate efficient retrieval and classification. Each question is assigned a unique index or identifier. A simple list or array structure is employed, where each element represents a question. The Fisher-Yates Shuffle algorithm is used to randomly permute the question indices. This ensures a random sequence for displaying the questions.

The algorithm is demonstrated as:

```
function fisherYatesShuffle ($array)
{
$count= count ($ array);
for ($i= $ count-1;$i>0;$i--)
{
//Generate a random index between 0 and $ i
$j=mt_rand (0, $i);
if ($i!=$j)
{
//Swap the elements at indices $i and $j
$temp = $ array [$i];
$array[$i] = $ array [$j];
$array [$j] = $temp;
}
}
Return $ array;
}
```

The function fisherYatesShuffle takes one parameter, $array, which is the array to be shuffled. The count function is used to determine the number of elements in the array and store it in the variable $count. A for loop is utilized to iterate from the last element (index $count - 1) to the second element (index 1). The loop variable $i starts at $count - 1 and decrements until it reaches 1. Within the loop, a random index $j is created using the mt_rand function. The random index is generated between 0 and $i.

An if statement checks if $i is not equal to $j. If they are not the same, the elements at indices $i and $j are swapped. A temporary variable $temp is used to hold the value of $array[$i]. Then, the value of $array[$j] is assigned to $array[$i], and the value stored in $temp is assigned to $array[$j]. This efficiently swaps the elements at indices $i and $j. After the loop completes, the shuffled array is returned.

The dataset used in this research is presented in a comma-separated value (CSV) format and includes the following columns: SN, Topic, Keyword, Question Patterns 1, Question Patterns 2, Question Patterns 3, and Responses. Figure 1 shows the proposed system architecture. The components and their interactions to create the intended system are depicted in the system architecture. It illustrates the procedure for adding question datasets to the data lake to enrich it. Preprocessing is performed to optimize and make the question datasets queryable. The randomization algorithm is utilized to generate a series of distinct inquiries.

The proposed system's architecture illustrates how users interact with the datasets in the data lake. The data lake serves as a repository for the imported datasets, resulting in the creation of a vast and centralized repository.

Preprocessing involves removing extraneous data from the dataset and ensuring that the question datasets are formatted meaningfully. Optimized question datasets are produced during the preprocessing phase.

To achieve the desired results, the dataset is queried, and the randomization algorithm is applied. The system processes the request and generates a list of unique questions based on the algorithm. Table 1 shows how the proposed computer science curriculum questions are designed using the data lake.

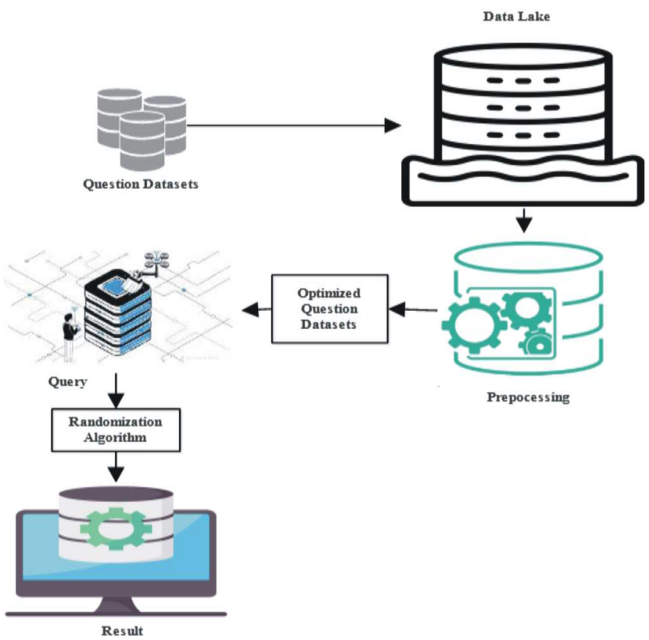

Figure 1. Proposed System Architecture

Input: data sets from ACM/IEEE computer science curriculum

Output: returns preview questions

Parameters: ACM/IEEE dataset (t), course (c), generate question (g_q), data lake (d), cleaned_questions(c_q), final_question(f_q)

Procedure

1. Input t
2. Process t g_q
3. For Each c
4. Store g_q into d
5. Query g_q
6. Preprocess g_q c_q
7. If g_q = c_q Then
8. Optimize c_q
9. Else
10. Goto 6
11. EndIf
12. Apply function FisherYatesShuffle ($array)
13. Generate f_q
14. Returns to 4
15. EndLoop

Table 1. Algorithmic Representation of the Proposed System

As shown in Table 1, lines 1 to 11 provide course synopses harvested from the ACM or IEEE computer science curriculum. The input results are processed to generate questions, and each question generated for a course is stored in a question bank referred to as the data lake. Line 5 queries the generated questions to preprocess them into cleaned questions. If the generated questions are cleaned, they are optimized; otherwise, the process loops back to the preprocessing stage.

Assuming all conditions are met, from line 12 onward, the algorithm is invoked to randomize the questions to generate the final cleaned questions. The loop continues until a specified threshold is reached. Furthermore, a sequence diagram is used to represent the proposed system. Figure 2 shows the sequence diagram for the proposed system.

The lecturer activates the system by entering login credentials. Once the correct account is identified, access is granted. Using the provided user interface, the lecturer configures the settings or options for generating questions. The lecturer then requests the system to generate questions. The system employs the randomization algorithm to perform this task. After processing the lecturer's request, the system generates a list of distinct questions. The lecturer concludes the session by logging off. The web technologies and stacks that are used to develop the system: Visual Studio Code, WAMP Server, Firefox, and Chrome browsers.

3. Results and Discussion

The development tools used to implement the system include the code editor Visual Studio Code and the development environment WAMP Server, which runs Apache, PHP, and MySQL services. This is done on a PC running 64-bit Windows 11 with 500 GB of RAM and an AMD Ryzen 7 4800H processor with 2.9 GHz and Radeon graphics. WAMP Server serves as the development environment for the system, while Apache, PHP, and MySQL services are used. Visual Studio Code is employed as the code editor.

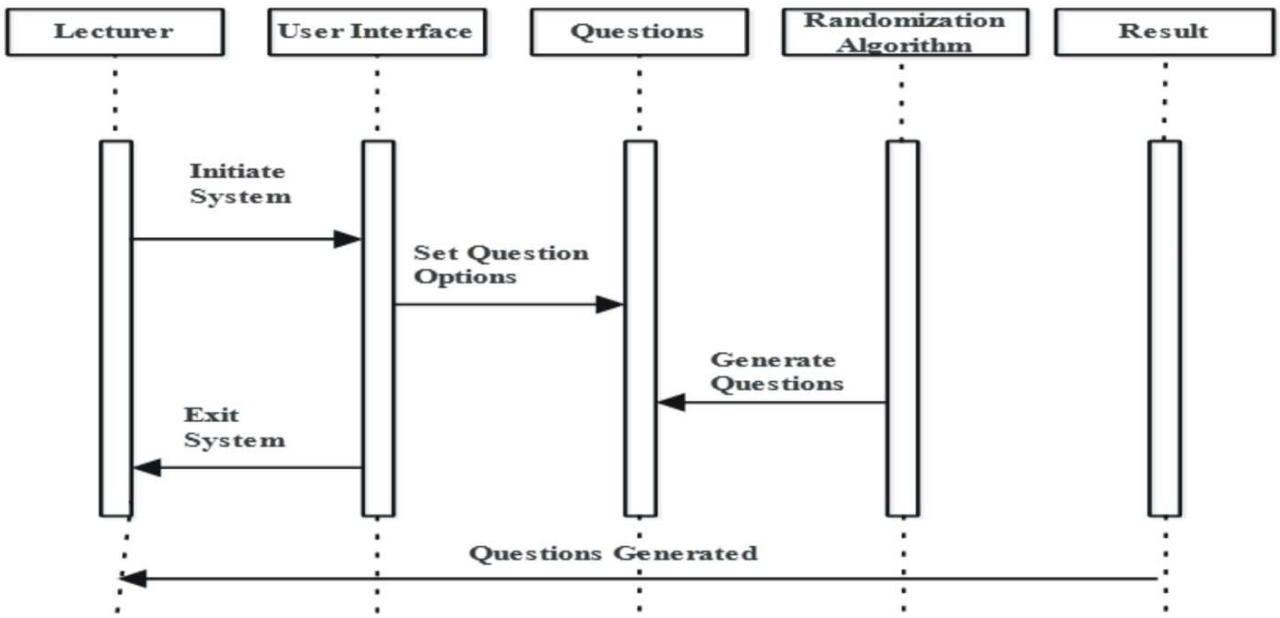

Figure 2. Sequence Diagram of the Proposed System

To store the created datasets in one place, a database is required. The database schema consists of four tables: admins, courses, questions, and topics, created using the MySQL database. Web technologies and stacks such as HTML5, CSS3, Bootstrap 5, and JavaScript are used to create responsive designs for user interfaces (UIs) and to render components dynamically. PHP is primarily used as the server-side technology for creating web-based applications that run on web servers, with features such as dashboards, inquiry pages, print logic, and login options.

The Fisher–Yates Shuffle method, which effectively shuffles a series of questions, is implemented in the system using PHP. The system provides user interfaces that incorporate forms, menus, and icons to facilitate user interaction. The user interface is intuitive, requiring only clicks or item selections. Figure 3 shows a login form with text input fields for the username and password and a control for the login button. If the login credentials are accurate, verified, and authenticated, the user is granted access to the system.

As shown in Figure 3, the user has the ability to input login details after completing the necessary registration to access the system. Figure 4 shows the interface of a question page for courses. Figure 5 shows the interface of a question page for topics. Similarly, Figures 4 and 5

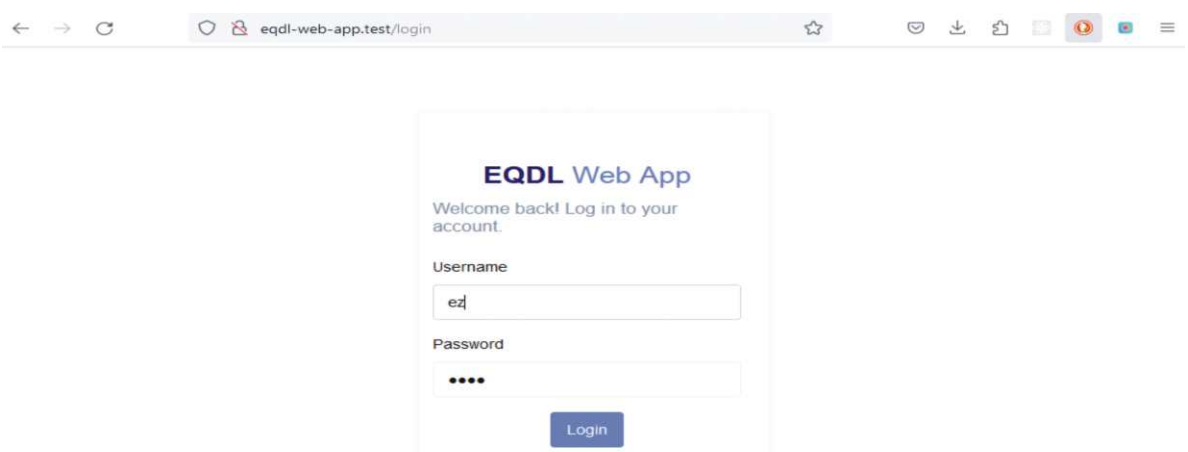

Figure 3. Login Page of the Proposed System

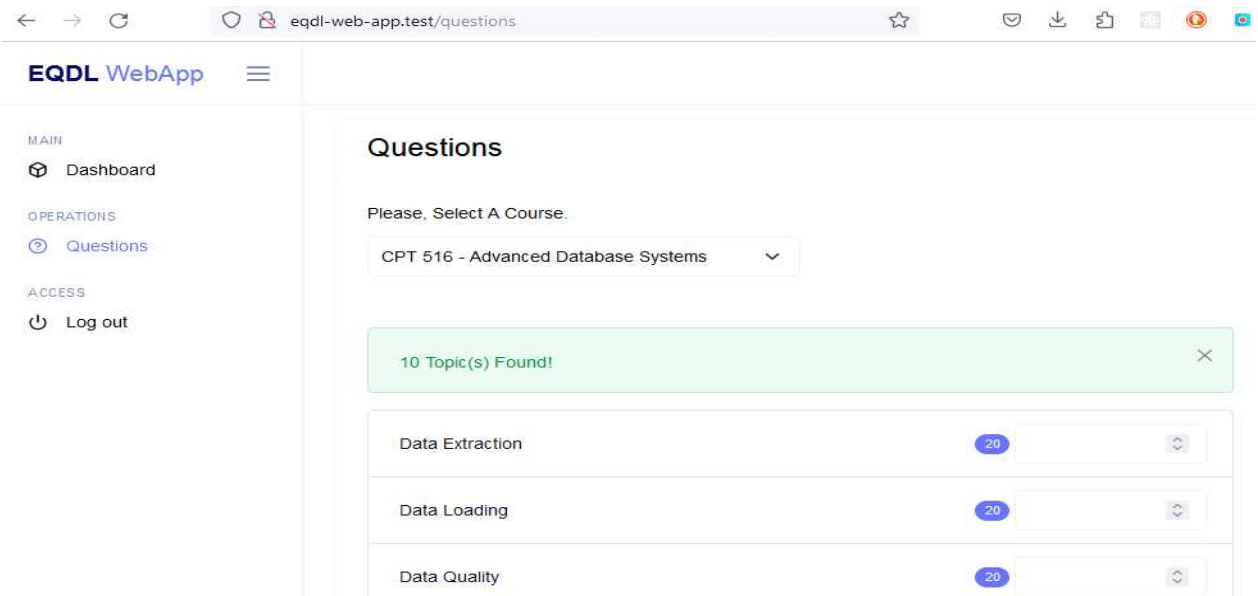

Figure 4. Interface of a Question Page for Course

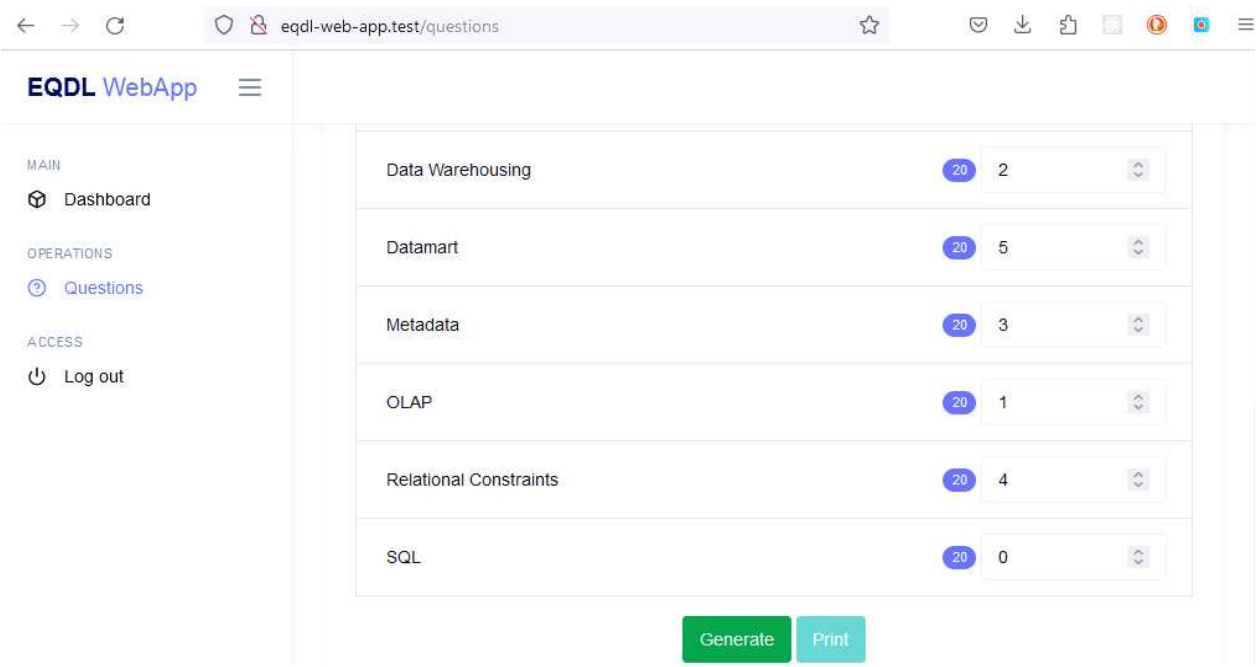

Figure 5. Interface of a Question Page for Topics

feature a dropdown form control with a selection of available courses. When a course is selected from the course menu, the related topics for that course are returned.

The course subjects are displayed on the interface, along with an indication of how many questions are available in each topic. There is also a number input field where users can select the number of questions they would like to generate. Once the desired number of questions is entered, the Generate button can be clicked. This initiates the validation process, which checks whether the number of questions entered falls within the permitted range. If

everything proceeds as expected, the questions can be printed by clicking the Print button. If there is an error, an error message will be displayed.

As shown in Figure 4, the page assists the user in navigating through questions by selecting a course. For example, a course coded CPT516 and titled "Advanced Database Systems" is inputted, and 10 topics are returned. More importantly, the statistics for data extraction, loading, and quality are provided. Consequently, Figure 5 shows the user to choose a topic for the questions to be generated and randomized. Based on the query displayed in Figure 5, a number of topics, including data warehouse, metadata, SQL, and others, were returned. The user can then trigger the generation command to produce questions based on the selected topic.

Figure 6 shows the "Print Question" page. The "Print Questions" page can be downloaded as a Portable Document Format (PDF) file or viewed directly in the web browser. The outcomes of the system's testing using datasets for the Advanced Database Systems (CPT516) course were distinct, randomly generated, and objective. The interface for Advanced Database Systems (CPT516) displays a total of fifteen (15) examination questions.

Figures 7 and 8 show the test cases for the login and Fisher-Yates shuffle algorithm. PHPUnit, a framework for writing and executing PHP code tests, was used for this. It ensures that specific code segments function as anticipated and behave appropriately under various circumstances.

Figure 7 shows the command view of the login test case for the server-side execution, displaying an execution time of 20 seconds.

Figure 8 shows the randomization algorithm test case. It shows a command view of the Fisher-Yates Shuffle as the randomization algorithm, displaying an execution time of 18 seconds. Table 2 shows unit testing using PHPUnit.

Table 2 shows the statistics in terms of execution time for the Login, Questions, Fisher-Yates Shuffle, Generate

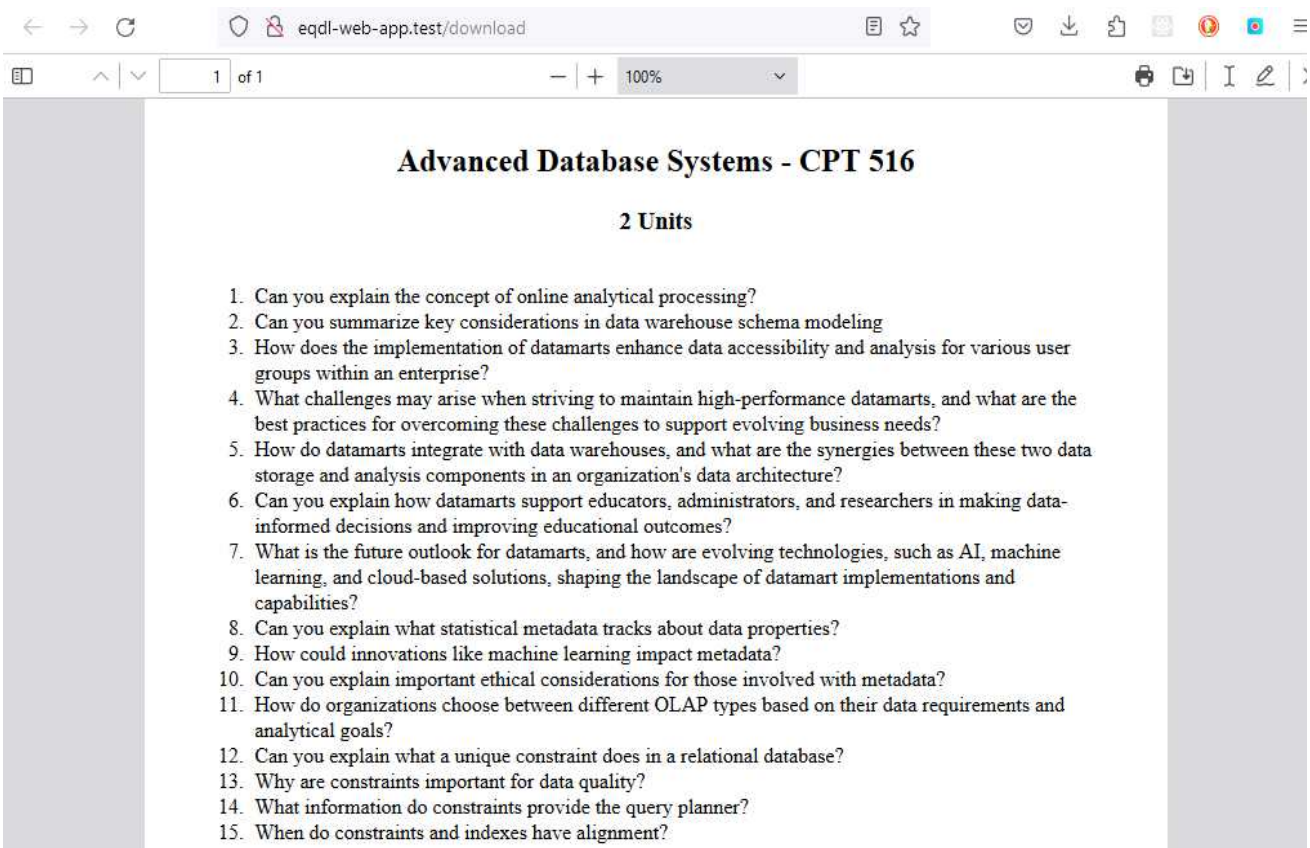

Figure 6. Interface of the Print Question Page

Figure 7. Command View of the Login Test Case

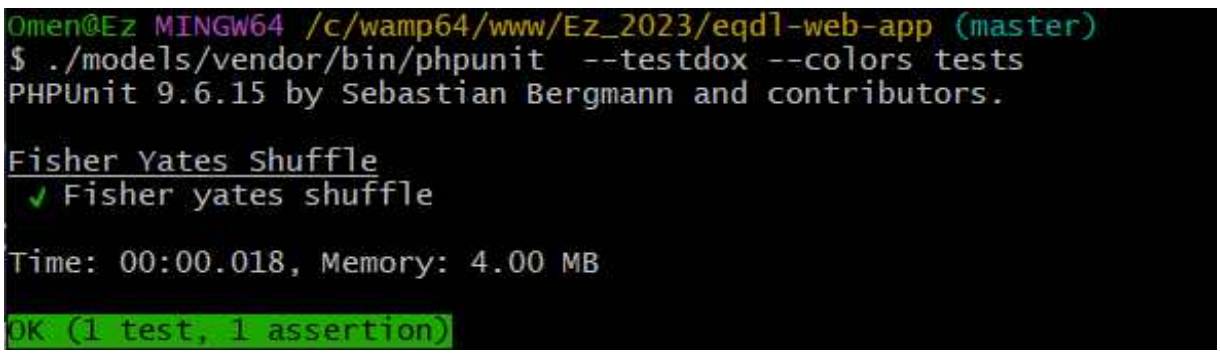

Figure 8. Command View of the Fisher-Yates Shuffle Test Case

SN	Test	Time	Memory	Passed
1.	Login	00:00.020	4.00MB	True
2.	Question	00:00.022	4.00MB	True
3.	Fisher-Yates Shuffle	00:00.018	4.00MB	True
4.	Generate Questions	00:00.016	6.00MB	True
5.	Courses List	00:00.012	4.00MB	True

Table 2. Unit Testing using PHPUnit

Questions, and Courses, which are 20, 22, 18, 16, and 12 seconds, respectively.

As a result of the impact and implementation of the proposed algorithmic design, the execution time for generating questions and listing courses has started to decline, which is a commendable improvement in this study. However, if other structured web-based technologies, such as XML, are employed, it might further improve the execution time.

Conclusion

In conclusion, this study designs a question data lake system with an essay-based structure in the computer science curriculum. Different data sources, such as ACM or IEEE computer science curricula, were considered for question formulation. The datasets created were imported into a database that was built using MySQL, and a web-based system with Graphical User Interfaces (GUIs) was developed to allow users to interact with the system. The system was developed using HTML5, CSS3, Bootstrap 5, PHP, and JavaScript. The Generative Artificial Intelligence and Fisher-Yates Shuffle algorithms were part of the methods considered to acquire the data sources for the questions.

The use of the Fisher-Yates Shuffle algorithm in an essay-based question data lake system augments the heterogeneity and stochasticity of the questions posed to users, guaranteeing a more equitable evaluation procedure. However, in this study, the scope of the curriculum considered is for the higher levels (especially the 500-level, final year) of the Department of Computer Science at the Federal University of Technology, Minna, Nigeria. This served as a case study to create original questions with straightforward and effective procedures.

After querying the data lake, a PDF file containing questions for a course is produced. This file can be downloaded and printed immediately.

These results show how educational institutions can use data lakes as a repository for inquiry, maximizing the potential of questions. Educational institutions can access the enormous amount of questions stored in the data lake and use them as a valuable resource for tasks such as research, analysis, and process improvement. It is advised that educational institutions and organizations use this method to control, monitor, and specify how different questions are used. Furthermore, by employing different techniques and methodologies on the dataset, this study suggests further investigation. To improve the data lake, additional datasets from different courses can be created. Improved web technologies, such as eXtensible Markup Language (XML), can be used as alternatives to HTML-based technologies. The Fisher-Yates algorithm can also be adapted to other disciplines, such as statistics and data analysis, where survey sampling is done by randomly selecting participants for surveys, ensuring unbiased and representative samples. It can also be used to shuffle and anonymize sensitive data. Studies can utilize this algorithm to randomize controlled trials, thereby reducing bias and ensuring reliable results.

Similarly, the proposed system is scalable to accommodate an extended or modified curriculum in the domain under consideration. Moreover, the system can be adapted to function optimally in other fields of study, thanks to the strategies and techniques employed in its development. Additionally, studies can employ semantic-based technologies to modify the data lake, ensuring that language ambiguities and mismatches in terms of synonyms and polysemy are adequately addressed.

References

[1]. Abdullah, N. F. (2023). Pengembangan dan pengujian aplikasi computer based test dengan menggunakan algoritme fisher-yates shuffle sebagai pengacakan soal. *Journal of Computer Science and Engineering (JCSE),* 4(1), 50-60.

[2]. Allam, A. M. N., & Haggag, M. H. (2012). The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS),* 2(3), 1-12.

[3]. Bhirangi, R., & Bhoir, S. (2016). Automated question paper generation system. *International Journal of Emerging Research in Management &Technology,* 5 (4), 28-34.

[4]. Clariana, R. B., & Solnosky, R. (2023). *Anchoring Concepts Influence Essay Conceptual Structure and Test Performance.* International Association for Development of the Information Society.

[5]. Errami, S. A., Hajji, H., El Kadi, K. A., & Badir, H. (2023). Spatial big data architecture: From data warehouses and data lakes to the Lakehouse. *Journal of Parallel and Distributed Computing,* 176, 70-79.

https://doi.org/10.1016/j.jpdc.2023.02.007

[6]. Essien, U. E. (2023). Design of Question Generator System (Qpgs) using fisher-yates shufling algorithm. *Technology,* 6(1), 66-82.

[7]. Hai, R., Koutras, C., Quix, C., & Jarke, M. (2023). Data lakes: A survey of functions and systems. *IEEE Transactions on Knowledge and Data Engineering,* 35(12), 12571-12590.

https://doi.org/10.1109/TKDE.2023.3270101

[8]. Hai, R., Quix, C., & Zhou, C. (2018). Query rewriting for heterogeneous data lakes. In *Advances in Databases and Information Systems: 22nd European Conference, ADBIS 2018, Budapest, Hungary, September 2–5, 2018, Proceedings 22* (pp. 35-49). Springer International Publishing.

https://doi.org/10.1007/978-3-319-98398-1_3

[9]. Haryanto, Y. (2023). Designing educational games with fisher yates shuffle method in early childhood at TKIT Al muqorrobin based on android. *Instal: Jurnal Komputer*, 15(02), 414-423.

https://doi.org/10.54209/jurnalkomputer.v15i02.165

[10]. Jiang, Z., Gao, T., Yin, Y., Liu, M., Yu, H., Cheng, Z., & Gu, Q. (2023, July). Improving domain generalization for prompt-aware essay scoring via disentangled

representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1, 12456-12470.

https://doi.org/10.18653/v1/2023.acl-long.696

[11]. Kumar, A. P., Nayak, A., K, M. S., Chaitanya, & Ghosh, K. (2024). A novel framework for the generation of multiple choice question stems using semantic and machine-learning techniques. *International Journal of Artificial Intelligence in Education,* 34(2), 332-375.

https://doi.org/10.1007/s40593-023-00333-6

[12]. Kuppusamy, P., & Suresh Joseph, K. (2022, April). Building an enterprise data lake for educational organizations for prediction analytics using deep learning. In *Proceedings of International Conference on Deep Learning, Computing and Intelligence: ICDCI 2021* (pp. 65-81). Springer Nature Singapore.

https://doi.org/10.1007/978-981-16-5652-1_6

[13]. Li, H., Li, C., Xing, W., Baral, S., & Heffernan, N. (2024, March). Automated feedback for student math responses based on multi-modality and fine-tuning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 763-770).

https://doi.org/10.1145/3636555.3636860

[14]. Lyu, J. (2022). Writing assistant scoring system for English second language learners based on machine learning. *Journal of Intelligent Systems,* 31(1), 271-288.

https://doi.org/10.1515/jisys-2022-0009

[15]. Nambiar, A., & Mundra, D. (2022). An overview of data warehouse and data lake in modern enterprise data management. *Big Data and Cognitive Computing,* 6(4), 132.

https://doi.org/10.3390/bdcc6040132

[16]. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research,* 21(140), 1-67.

[17]. Risnasari, M., Effindi, M. A., Dellia, P., Cahyani, L., & Aini, N. (2021). Computer based test using the fisher-yates shuffle and smith waterman algorithm. *KnE Social Sciences* (pp. 353-360).

https://doi.org/10.18502/kss.v5i6.9224

[18]. Shabani, N., Beheshti, A., Farhood, H., Bower, M., Garrett, M., & Alinejad-Rokny, H. (2023). A rule-based approach for mining creative thinking patterns from big educational data. *AppliedMath,* 3(1), 243-267.

https://doi.org/10.3390/appliedmath3010014

[19]. Singhal, R., Goyal, S., & Henz, M. (2016, November). User-defined difficulty levels for automated question generation. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 828-835). IEEE.

https://doi.org/10.1109/ICTAI.2016.0129

[20]. Sultan, M. A., Chandel, S., Astudillo, R. F., & Castelli, V. (2020, July). On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5651-5656).

https://doi.org/10.18653/v1/2020.acl-main.500

[21]. Thigale, P., Nadkar, S., Kokle, B., Bhole, N., & Ghodichor, F. S. (2021). Automatic question paper generator system by keyword based shuffling algorithm using randomization technique. *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE),* 10, 573-577.

[22]. Wang, X., Wang, Z., Huang, W., Wen, G., & Zhang, S. (2017, May). Design and realization of test question bank database system. In *Materials Science, Energy Technology, and Power Engineering i: 1st International Conference on Materials Science, Energy Technology, Power Engineering (MEP 2017),* 1839 (1), 020196.

https://doi.org/10.1063/1.4982561

[23]. Zaib, M., Zhang, W. E., Sheng, Q. Z., Mahmood, A., & Zhang, Y. (2022). Conversational question answering: A survey. *Knowledge and Information Systems,* 64(12), 3151-3195.

https://doi.org/10.1007/s10115-022-01744-y

[24]. Zhang, J., & Takuma, J. (2015, October). A Kanji learning system based on automatic question sentence generation. In *2015 International Conference on Asian Language Processing (IALP)* (pp. 144-147). IEEE.

https://doi.org/10.1109/IALP.2015.7451552

ABOUT THE AUTHORS

Dr. Enesi Femi Aminu is presently a Senior Lecturer in the Department of Computer Science, School of Information and Communication Technology, Federal University of Technology (FUT), Minna, Nigeria. He has taught various courses, including Artificial Intelligence, Expert Systems, Operating Systems, and Advanced Database Design. He obtained his Ph.D. in Computer Science from FUT Minna, Nigeria. Similarly, he earned both his B.Sc. and M.Sc. degrees in Computer Science from the University of Jos, Jos, and Ahmadu Bello University, Zaria, Nigeria, respectively. His research interests include Knowledge Representation, specifically Ontology Design and Semantic Search, Artificial Intelligence, and Machine Learning. He is also a member of several professional bodies, including the Computer Professionals Registration Council of Nigeria (CPN), the Nigeria Computer Society (NCS), and the International Association of Computer Science and Information Technology (IACSIT). He has successfully published over 35 peer-reviewed scientific articles in reputable outlets, including journals, international conferences, and book chapters.

Dr. Oluwaseun Adeniyi Ojerinde is a Senior Lecturer in the Department of Computer Science in the School of Information and Computer Technology at the Federal University of Technology, Minna, Nigeria. He is an expert in mobile communications and computer science with a broad range of experience in blockchain technology, digital transformation and innovation, process optimization, antenna systems, and radio propagation. He bagged his B.Sc. in Computer Technology at Babcock University, Nigeria, in 2006. He received his M.Sc. in Mobile Communication Systems from Loughborough University, Nigeria, in 2008. He also obtained his PhD in Mobile Communication Systems from Loughborough University, Nigeria, in 2014. His research area is encryption and Blockchain Technology for Data Management, Process Optimization, 5G Technology, Antenna Systems and Propagation, On-Body Systems, Multiple Input Multiple Output (MIMO) Systems, Telecommunications, Specific Absorption Rate (SAR), Radiation, and Computer Science Education. He is a member of CPN and IEEE.

Ayobami Ekundayo is a Lecturer in the Department of Computer Science at the School of Information and Communication Technology, Federal University of Technology Minna, Niger, Nigeria. She is running her PhD at the Federal University of Technology Minna, Niger, Nigeria. She received both B.Sc. and M.Sc. Computer Science from the University of Ilorin, Kwara, Nigeria. Her research interests include Data Mining. She is a member of CPN (Computer Professionals Registration Council of Nigeria). She has published some research papers in reputable journals and conferences.

Ezra Adamu was an Undergraduate Student of the Department of Computer Science, School of Information and Communication Technology, Federal University of Technology Minna, Niger State, Nigeria. He graduated in 2023, and this publication is an excerpt of his undergraduate project. He admires the application of Data Lake in the area of any educational institute. He proposed to further his research interests in Algorithms and Methods to enrich the Data Lake.

Abisoye Opeyemi Aderlike is an Associate Professor of Computer Science and the Head of the Department (since 2021) at the Computer Science Department, Federal University of Technology, Minna, Niger State, Nigeria. She holds B.Sc., M.Sc., and Ph.D. degrees in Computer Science from the University of Ilorin, Nigeria. She is a registered member of the Computer Professionals Registration Council of Nigeria (MCPN) and the Nigeria Computer Society (NCS). Her research interests focus on Artificial Intelligence, Software Engineering, High-Performance Computing, and Bioinformatics. She has authored and co-authored over 40 scholarly articles published in reputable national and international journals and conferences.

Hussaini Abubakar Zubairu is a Lecturer in the Department of Information Technology at the Federal University of Technology, Minna, Nigeria. He is currently pursuing a Ph.D. at Carleton University, Canada. He holds a B.Sc. and M.Sc. in Computer Science from Ahmadu Bello University, Zaria, Nigeria. His research interests include Machine Learning, Misinformation Detection and Intervention, and Knowledge Management. He has published research papers in reputable journals, making significant contributions to his fields of interest.

Lawal Olamilekan Lawal is a Lecturer in the Department of Computer Science, School of Information and Communication Technology, Federal University of Technology, Minna, Niger State, Nigeria. He is currently pursuing his M.Sc. at Kwara State University, Malete, Kwara State, Nigeria. He earned his B.Sc. in Computer Science from Al-Hikmah University, Ilorin, Kwara State, Nigeria. His research interests include Computer Vision. He has presented a paper at a conference.

i-manager® Publications

3/343, Hill view, Town Railway Nager, Nagercoil
Kanyakumari Dist. Pin-629 001.
Tel: +91-4652-231675, 232675, 276675

e-mail: info@imanagerpublications.com
contact@imanagerpublications.com
www.imanagerpublications.com

i-manager Printing Press