

Automated ER Diagram Generation, Plagiarism Check and AI-Based Quiz System

Wanniarachchi M.Y, Kapukotuwa S.A.A.H, Kumarasinghe C.A.W, De Silva S.K.S,
Samantha Rajapaksha, Pasangi Ratnayake

Dept. of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

Abstract: - The proposed study covers a range of topics, including entity-relationship (ER) diagram analysis and extraction, plagiarism detection in ER diagrams, AI-based quiz systems, and database architecture and administration. The first section focuses on applying natural language processing to automate ER diagram development and SQL query formulation, possibly reducing time, and assisting non-technical users in database access and decision-making. The integrity of database systems is an issue; thus the second section provides a plagiarism detection component that contrasts visual and quantitative accuracy. The final section introduces an AI-based quiz system that fills in the gaps left by traditional quizzing techniques by providing individualized feedback, dynamically produced questions, and the capacity to assess text-based solutions. The last section describes the creation of a machine-learning model for precise entity recognition and attribute extraction from ER diagrams, offering a useful tool for academics. While filling in specific research gaps, this extensive study advances several fields, including database systems, AI-based teaching, and ER diagram analysis.

Keywords: *Automated, Entity-Relationship Diagrams, Query Generation, Natural Language Processing, Database Management, Machine Learning, SQL Queries.*

1. Introduction

The subjectivity and inefficiency of the present approaches for grading student work in Entity-Relationship (ER) diagrams are addressed in this study. In disciplines including computer science, business management, and engineering, these diagrams are often utilized; nevertheless, manual assessment may be labor-intensive and arbitrary. It might be challenging for professors to recognize instances of plagiarism and unlawful cooperation and to impose the required sanctions. Additionally, the methodologies used today might not truly represent how well pupils comprehend the subject. Based on the students' ER diagrams and inquiries, an automated system may produce ER diagrams, check for plagiarism, and suggest questions. In many different professions, this method can increase fairness, effectiveness, and academic integrity.

A. Automated Er Diagram Generation AND Query Generation

Entity-relationship (ER) diagrams are critical for developing and comprehending database systems in the present digital era, when data is essential for a company's success. But for non-technical people, producing SQL queries and ER diagrams might be difficult. Systems that automatically generate SQL queries and ER diagrams from input in plain English are currently being developed. The system generates suitable ER diagrams and database structures using machine learning and natural language processing (NLP) approaches, facilitating access to and manipulation of data contained in databases. The major objective of the system is to make data retrieval and database architecture simpler for non-technical users.

Novel features can improve database architecture and query creation, including automated attribute recognition, contextual information inclusion, weak entity identification, table joins, and nested queries. By taking a fresh perspective to database architecture and data retrieval issues, the suggested research component might potentially reduce user engagement, save time and money, and get rid of mistakes and inconsistencies. To evaluate the effectiveness and usability of the system in real-world circumstances, more research is required.

B. ER Diagram Analysis and Extraction

By utilizing machine learning models and data processing techniques, the proposed automated ER diagram analysis and evaluation component seeks to increase the precision and effectiveness of ER diagram analysis and assessment. This component analyzes the validity and completeness of extracted entities, properties, and relationships from ER diagrams using SVM and NLP. The element simplifies the procedure, lessens the effort for the evaluators, and enhances the precision and consistency of evaluations. In order to evaluate the accuracy and completeness of the entities, characteristics, and connections, the component compares the ER diagrams created by students and evaluators. An evaluation method gives points for accurate identification and extraction while taking points away for either. In general, the automated ER diagram analysis and evaluation component presented increases the accuracy and efficiency of ER diagram analysis and assessment while streamlining the process.

C. Comparison of Visual and Quantitative Similarity

Although entity-relationship (ER) diagrams are a common tool for database systems, an effective plagiarism detection tool has been developed in response to the rising concern about plagiarism in academic and professional contexts. In order to assess the visual and quantitative accuracy of ER diagrams, as well as the handwritten typefaces and diagram dimensions, this research project intends to develop a plagiarism detection program. This tool will be helpful in academic and professional situations where originality is crucial and plagiarism is becoming a bigger problem.

The research will include a review of the literature to determine the visual and quantitative similarity of ER diagrams, gathering a large collection of ER diagrams from academic and professional contexts, developing metrics to compare the visual and quantitative accuracy of ER diagrams, developing algorithms and methods for doing so, and putting the algorithms and techniques into a software program to evaluate the results.

The suggested instrument for detecting plagiarism will be more accurate than current instruments at doing so while maintaining the integrity of the original work. The tool will be more beneficial in academic and professional situations where uniqueness is important by evaluating the visual and quantitative correctness of ER diagrams, including the ability to compare handwritten typefaces and diagram dimensions.

D. AI-Base Quiz System

A contemporary approach to evaluating student learning and offering individualized feedback is to use quiz systems powered by AI. These systems examine text-based replies using Natural Language Processing (NLP) algorithms, recognizing keywords and phrases, and contrasting them with the correct answer(s). Based on the results of the analysis, the next question is selected, and unique feedback is generated for each learner. AI-based quiz systems are designed to give students an interactive and interesting learning experience while improving their abilities to engage in Self-Regulated Learning (SRL). AI-based quiz systems assist students in building these abilities and improving their learning through the provision of individualized feedback and dynamically produced questions. Overall, AI-based quiz systems present a potential strategy for improving students' SRL abilities and delivering a fun and dynamic learning environment.

2. Literature Review

A. Automated Er Diagram Generation AND Query Generation

Entity-relationship (ER) modeling has experienced significant advancements in recent years, with researchers exploring the use of natural language processing (NLP) techniques to generate ER diagrams from requirement specifications. Hettiarachchi et al [1] proposed a scenario-based ER diagram and query generation engine, leveraging NLP techniques to extract relevant information from requirement specifications and automatically generate ER diagrams. Omar et al [2] introduced heuristics-based entity-relationship modeling through NLP,

focusing on identifying relationships between entities based on the syntactic structure of natural language sentences. Btoush and Hammad [3] explored the generation of ER diagrams from requirement specifications using NLP, proposing a framework that applies various NLP techniques, including part-of-speech tagging and dependency parsing. Kashmira and Sumathipala [4] presented a methodology for generating ER diagrams from requirement specifications based on NLP, employing keyword extraction and semantic role labeling techniques to identify entities and relationships from natural language texts. These sentences and the importance of capturing relevant information, understanding the syntactic and semantic structures of natural language sentences, and employing heuristics and automated techniques for effective ER diagram generation.

B. ER Diagram Analysis and Extraction

Plagiarism detection and entity-relationship diagram (ERD) generation are crucial areas of research in information processing and software engineering. These papers focus on detecting plagiarism in images, analyzing image content to identify potential cases, and generating ER diagrams from system requirement specifications. Ovhall [5] proposes a method for detecting plagiarism in images, while Uduwela and Wijayarathna [6] conduct a survey on tools and systems for ER diagram generation. Kashmira and Sumathipala [7] propose a method for automatically generating ER diagrams using natural language processing (NLP) techniques, leveraging semantic analysis of textual requirements to extract entities and relationships. Dahanayake et al. [8] introduce a plagiarism detection tool for enhanced entity-relationship diagrams, incorporating syntactic and semantic analysis to identify instances of plagiarism. These papers collectively contribute to the fields of plagiarism detection and ER diagram generation by addressing specific challenges and proposing novel algorithms and techniques to improve accuracy and efficiency.

C. Comparison of Visual and Quantitative Similarity

In recent years, plagiarism has become a significant concern in academic research and education. Researchers have extensively studied various approaches and algorithms to detect and prevent plagiarism in textual documents. Nurdiansyah et al. [9] proposed the Winnowing Algorithm based on K-gram to identify instances of plagiarism in file text-based documents. Khosravi et al. [10] explored the application of Explainable Artificial Intelligence (XAI) in education, emphasizing the importance of AI algorithms providing clear explanations for decisions and actions, enhancing transparency and trust between users and AI systems. Foltýnek et al. [11] conducted a systematic literature review on academic plagiarism detection, examining various approaches, algorithms, and techniques employed in the field.

Hammad et al. [12] conducted a systematic mapping study of clone visualization, exploring techniques for identifying code clones in software development. These research papers contribute to the body of knowledge on plagiarism detection, explainable AI in education, and clone visualization, providing valuable insights for researchers and practitioners working in these domains.

D. AI-Base Quiz System

In recent years, there has been a growing interest in developing automated systems for generating exam questions and providing personalized quizzes. Wang et al. [13] introduced the IFSE system, a personalized quiz generator and intelligent knowledge recommendation platform, using semantic computing techniques to generate quizzes tailored to individual users' knowledge levels and preferences. Rahim et al. [14] presented a novel approach for generating exam questions using a genetic algorithm, optimizing the generation process by evolving a population of candidate questions. Immanuel and Tilasi [15] proposed a framework for automatic examination paper generation, focusing on well-structured question papers covering a wide range of topics. Butler and Winne [16] conducted a theoretical synthesis on feedback and self-regulated learning, highlighting the importance of providing effective feedback to learners to support their self-regulated learning processes.

Panday and Rajeswari [17] developed a system for automatic question generation using software agents specifically tailored for technical institutions, leveraging software agents to generate questions based on predefined rules and patterns. These advancements hold significant potential in revolutionizing educational assessment practices and enhancing students' learning experiences.

3. Methodology

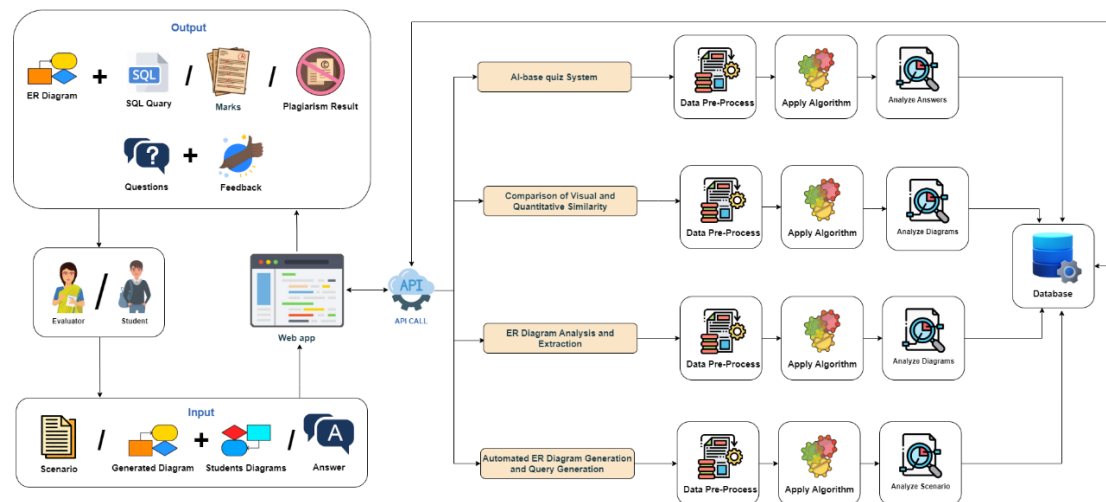


Figure 1. System Overview Diagram

A. Automated Er Diagram Generation AND Query Generation

The process for creating Entity-Relationship (ER) diagrams has changed, and it now makes use of the Logistic Regression model. To ensure alignment with the Logistic Regression model, this new technique starts with a thorough analysis of the dataset that includes data interpretation, management of missing values, encoding of categorical variables, and normalization of numerical properties. Entity connections are shown visually in the resultant ER diagram.

A Logistic Regression classifier is now used to draw the ER diagram. The model is trained using a dataset of items and the relationships that connect them, with labels given to the entities and categories given to the relationships. Based on this dataset, relationships between items are predicted; these relationships are depicted graphically by connecting lines. Then, to make the ER diagram more understandable, annotations are added to it to indicate the expected types of relationships. The Logistic Regression model's prediction skills as well as its capacity to identify patterns and correlations within the dataset are both taken advantage of by this improved technique.

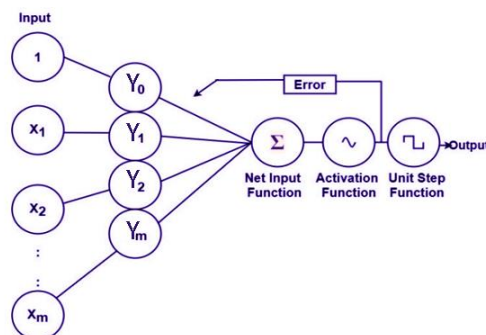


Figure 2. Logistic Regression Architecture

The ER diagram itself continues to provide a thorough representation of the dataset, including attribute data for each item. To represent these characteristics, matching rectangles are joined to corresponding oval nodes. In order to forecast connections, the Logistic Regression classifier takes center stage, while diligent evaluation and changes guarantee the correctness and completeness of the ER diagram.

By leveraging the Logistic Regression model to depict entity connections visually, this revised technique alters the creation of ER diagrams and eventually improves understanding of the underlying data structure.

B. ER Diagram Analysis and Extraction

The OCR model is utilized to extract vital data from Entity-Relationship (ER) diagrams, a crucial tool in database architecture. Data gathering involves preparing a diverse dataset of ER diagrams, representing various complexity levels and aesthetics. These diagrams are then preprocessed to enhance text readability before being used in the OCR model. This process ensures optimal performance and accuracy in extracting entities, properties, and connections from ER diagrams, contributing to the overall information extraction process.

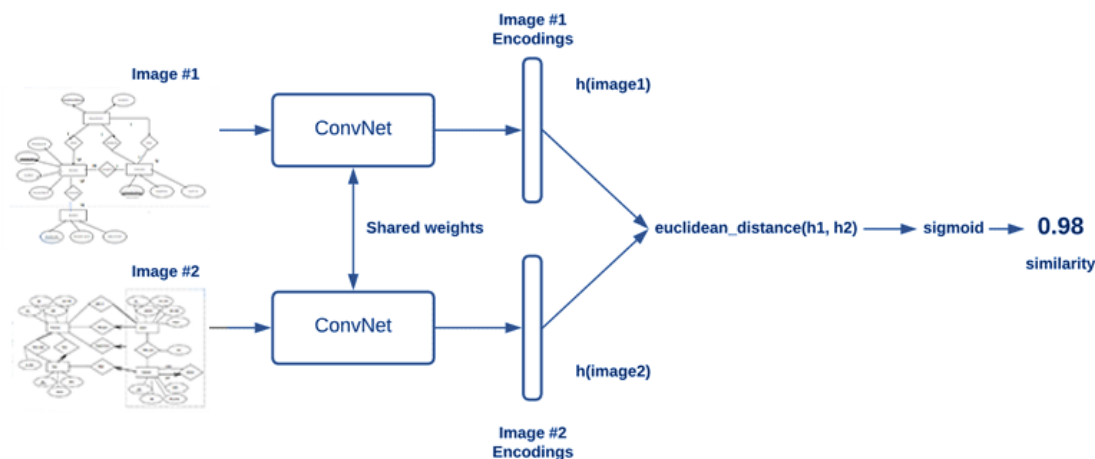


Figure 3. Keras OCR Architecture

The ER diagrams are processed using a deep learning OCR model built on the Keras framework. This model, trained on a large corpus of text images, accurately detects and transcribes text. Convolutional neural networks (CNNs) are used for feature extraction, while recurrent neural networks (RNNs) are used for sequence prediction. The model also recognizes entities, attributes, and relationships using a multi-step approach. Tokenization and syntactic analysis are used to identify relevant terms and language patterns, while rule-based techniques and semantic parsing are employed to categorize text segments into entities, properties, and relationships.

The ER diagrams undergo a post-processing stage to refine the findings and correct any errors made during previous processes. This process involves a validation procedure using manually labeled ER diagrams to assess the precision and efficacy of the proposed technique. The automated extraction of entities, characteristics, and connections from ER diagrams is made possible by a robust OCR model and a combination of linguistic and rule-based methodologies, thereby expediting database design and schema production.

C. Comparison of Visual and Quantitative Similarity

In this paper, we demonstrate a unique approach to plagiarism detection in Entity-Relationship (ER) diagrams by combining deep learning with visual feature extraction. For improved accuracy in spotting probable plagiarism, our method uses a thorough procedure that combines visual and quantitative analysis. The process starts with a selected dataset of ER diagrams, which are then scaled and translated into universal picture formats to maintain consistency. To extract high-level characteristics from the ER diagram pictures, we use the VGG16 model, a pre-trained convolutional neural network recognized for its ability to extract visual features.

By measuring the distance between feature representations of ER diagrams using well-established metrics like cosine similarity or Euclidean distance, visual similarity is assessed. Cardinality, connections, entities, and other quantitative properties are also retrieved and converted into numerical vectors for quantitative comparison. We detect possibly copied ER diagram pairings by setting a predetermined plagiarism threshold for both visual and quantitative similarity scores.

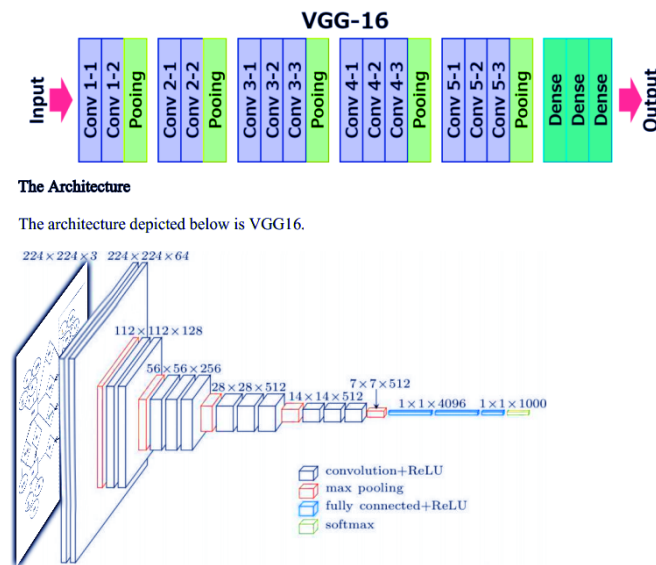


Figure 4. VGG-16 Architecture

Pairwise analysis is used in the comparison process, and ER diagrams that are visually or quantitatively more similar than the threshold are identified as possible instances of plagiarism. Manual evaluation of these identified instances is done to separate instances of academic dishonesty from true parallels in order to corroborate our findings. The technique includes a thorough reporting phase where discovered occurrences, similarity ratings, and any taken actions are painstakingly documented. Our method not only improves the precision of plagiarism detection in ER diagrams but also offers researchers and educators an organized and thorough framework for maintaining academic integrity in the field of database modeling and design.

D. AI-Base Quiz System

The development of an intelligent educational system, which autonomously generates questions, evaluates student responses, and provides personalized feedback, is detailed in this section. Central to this proposed system is the utilization of the Nearest Neighbors model, a powerful machine learning approach that enhances learning outcomes by tailoring educational content to the unique needs of individual students.

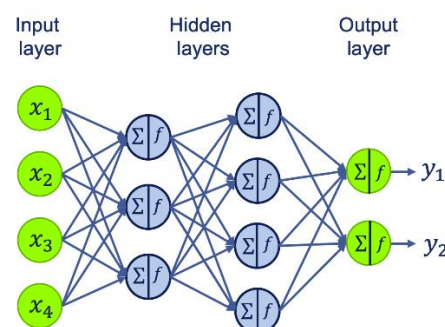


Figure 5. k-Nearest Neighbor (k-NN) Classifier Architecture

The initial phase involves comprehensive preprocessing of a vast dataset comprising questions and responses across diverse academic subjects. Employing contemporary techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings, the textual data undergoes vectorization, transforming it into numerical representations suitable for machine learning algorithms.

The cornerstone of the suggested system is the Nearest Neighbors model, a robust algorithm made accessible through the scikit-learn package. Through training on the vectorized question data, this model gains the ability to recognize semantically related questions. Leveraging this capability, the system can generate relevant questions based on specific topics.

Upon posing a question to the student, the system assesses their response by measuring the similarity between the student's answer and the correct response, utilizing cosine similarity as the distance metric. If the student's response aligns with the correct answer, the generated similarity score is compared to a predetermined threshold. If surpassing the cutoff, the response is deemed accurate, and the algorithm proceeds to formulate the next query. In cases of inaccuracies, the system provides tailored feedback and proposes follow-up questions to reinforce understanding of the particular topic.

Throughout this interactive exchange, the system maintains a record of the student's accurate responses, enabling the calculation of cumulative marks. These cumulative scores serve as a metric to evaluate the student's performance across various disciplines. The algorithm identifies areas where the learner may require additional assistance, offering targeted suggestions for further study based on their performance.

In conclusion, our innovative approach harnesses the power of the Nearest Neighbors model to construct an adaptive educational system. This system seamlessly generates questions, evaluates responses, and delivers personalized feedback, with the overarching goal of enhancing student engagement and fostering successful learning outcomes through the integration of machine learning techniques into the educational process.

Table 1. CLASSIFICATION MODEL PERFORMANCE SUMMARY

	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.83	0.71	0.99	0.83
Keras OCR	0.92	0.89	0.94	0.91
VGG-16	0.78	0.85	0.73	0.79
k-Nearest Neighbor	0.88	0.82	0.91	0.86

4. Results and Discussion

Our revolutionary research initiative stands as a formidable accomplishment in the realms of education and database management. Harnessing the prowess of a robust and cutting-edge technology stack, we have meticulously crafted a web-based educational system that caters to the needs of both educators and students alike.

At the heart of this innovation lies the backend, an intricately designed system implemented in Python. This not only ensures a seamless and scalable experience but also provides the flexibility required to adapt to the ever-evolving landscape of educational technologies. To facilitate efficient communication, we've incorporated a Flask-based Restful API, serving as the backbone that propels the system's functionality.

On the frontend, an aesthetically pleasing and user-friendly interface takes center stage. Crafted with a blend of HTML, CSS, and JavaScript, the interface is further enriched by the inclusion of Bootstrap, ensuring not only responsiveness but also an engaging visual appeal that enhances the overall user experience.

One of the standout features of our project is the Automated ER Diagram and Query Generation system. This groundbreaking component allows evaluators and students to input a scenario, triggering the automatic generation

of the corresponding Entity-Relationship (ER) diagram and SQL queries. This not only simplifies the understanding of database schemas but also significantly streamlines the educational and evaluative processes.

Adding another layer of sophistication is the ER Diagram Analysis and Extraction component. This empowers evaluators to efficiently compare student-generated ER diagrams with a reference diagram, facilitating a streamlined assessment process focused on correctness and completeness.

Venturing beyond conventional plagiarism detection, our research introduces a groundbreaking feature—the Comparison of Visual and Quantitative Similarity. This feature enables evaluators to assess not only the visual correctness but also the quantitative aspects of ER diagrams. This includes the calculation of similarity percentages for hand-drawn diagrams, ensuring the utmost integrity in the work submitted by students.

Complementing these components is the incorporation of an AI-based Quiz System, a pinnacle of innovation in educational technology. Educators can effortlessly customize quizzes, specifying topics and question numbers, while students access and answer questions with the assurance of automated evaluation and instant feedback. This not only simplifies the quiz creation process for educators but also enriches the learning experience for students.

In conclusion, our research project stands as a beacon of transformation in the educational landscape of database management. The fusion of a powerful technology stack with innovative components establishes a dynamic and enriching learning environment. The automated ER diagram processes, coupled with the AI-based quiz system, redefine the educational experience, ensuring educators and students have access to a comprehensive and highly effective platform.

5. Conclusion

This paper introduces a heuristic-based technique for automated entity, attribute, and relationship extraction, facilitating ER diagram creation and translation of natural language queries into SQL. The emphasis on proper English grammar is crucial, as non-English languages are currently unsupported. The system lacks user context awareness, warranting further investigation.

Highlighting the heuristic technique's impressive precision, the study advocates for increased accuracy through the discovery of new heuristics. The method's potential expansion includes identifying weak entities, assessing relationship strength, and providing Enhanced ER diagrams. Additionally, the SQL query module may incorporate table joins and nested queries, enhancing usability for non-technical users and enabling data retrieval through plain language queries.

Addressing picture segmentation challenges, future research aims to improve model effectiveness and explores in-memory database systems for accelerated data storage and reduced I/O operations. In the educational realm, NLP-driven analysis holds promise for providing fact-based questions or recommendations to teachers, utilizing diverse student learning data for machine learning-driven evaluations.

The extensive roadmap outlines goals such as refining heuristic precision, broadening application, enhancing SQL query capabilities, expediting picture segmentation models, creating in-memory databases, leveraging NLP for teaching support, and advancing learning analytics. Together, these trajectories propel the system's effectiveness, usability, and adaptability, ushering in a revolutionary era for database administration and instructional facilitation.

In conclusion, the research spans ER diagram plagiarism detection, AI-based quiz systems, and ER diagram analysis and extraction. The automated component enhances non-technical users' database access, streamlining procedures. Machine learning-driven ER diagram analysis improves review speed and accuracy, while plagiarism detection addresses integrity issues. The AI-based quiz system promotes self-regulated learning.

Acknowledging drawbacks, future research may focus on a uniform ER diagram assessment framework and enhancing plagiarism detection precision with NLP. Expanding automatic review to other diagram types and improving user experience/interface design can further enhance the tool's usefulness. In summary, the research advances ER diagram analysis, AI-based teaching, and database systems, offering precise solutions for future

developments in quiz systems, plagiarism detection, and database architecture, ultimately enhancing academic and professional experiences.

References

- [1] S. Hettiarachchi, C. Sugandhika, A. Kathriarachchi, S. Ahangama and G. T. Weerasuriya, "A Scenario-based ER Diagram and Query Generation Engine," in 4th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 2019.
- [2] N. Omar, J.R.P. Hanna, P. McKeivitt, "Heuristics-based entity-relationship modeling through natural language processing," in Proc. of the 15th Artificial Intelligence and Cognitive Science Conference (AICS-04), Dublin, Ireland, 2004.
- [3] E. S. Btoush, M. M. Hammad, "Generating ER Diagrams from Requirement Specifications Based on Natural Language Processing," in International Journal of Database Theory and Application, 2015.
- [4] P. G. T. H. Kashmira and S. Sumathipala, "Generating Entity Relationship Diagram from Requirement Specification based on NLP," in 2018 3rd International Conference on Information Technology Research (ICITR), 2018.
- [5] P. Ovhal, "Detecting plagiarism in images," 2015 International Conference on Information Processing (ICIP), Pune, India, 2015, pp. 85-89, doi: 10.1109/INFOP.2015.7489356.
- [6] W. C. Uduwela and G. Wijayarathna, "Survey on tools and systems to generate ER diagram from system requirement specification," 2014 IEEE International Conference on Industrial Engineering and Engineering Management, Selangor, Malaysia, 2014, pp. 370-373, doi: 10.1109/IEEM.2014.7058662.
- [7] P. G. T. H. Kashmira and S. Sumathipala, "Generating Entity Relationship Diagram from Requirement Specification based on NLP," 2018 3rd International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 2018, pp. 1-4, doi: 10.1109/ICITR.2018.8736146.
- [8] H. Dahanayake, D. Samarajeewa, A. Jayathilake, D. Bandara, A. Karunasena and L. Weerasinghe, "Plagiarism Detection Tool for Enhanced Entity-Relationship Diagrams," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 2021, pp. 0598-0606, doi: 10.1109/UEMCON53757.2021.9666552.
- [9] Y. Nurdiansyah, F. Nur Muharrom, and F. Firdaus, "Implementation of Winnowing Algorithm Based K-Gram to Identify Plagiarism on File Text-Based Document," MATEC Web Conf., vol. 164, Apr. 2018, doi: 10.1051/MATECONF/201816401048.
- [10] H. Khosravi et al., "Explainable Artificial Intelligence in education," Comput. Educ. Artif. Intell., vol. 3, p. 100074, Jan. 2022, doi: 10.1016/J.CAEAI.2022.100074.
- [11] T. Foltýnek, N. Meuschke, and B. Gipp, "Academic plagiarism detection: A systematic literature review," ACM Comput. Surv., vol. 52, no. 6, Oct. 2019, doi: 10.1145/3345317.
- [12] M. Hammad, H. A. Basit, S. Jarzabek, and R. Koschke, "A systematic mapping study of clone visualization," Comput. Sci. Rev., vol. 37, p. 100266, Aug. 2020, doi: 10.1016/J.COSREV.2020.100266.
- [13] X. Wang et al., "IFSE - Personalized Quiz Generator and Intelligent Knowledge Recommendation," 2022 IEEE 16th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2022, pp. 201-208, doi: 10.1109/ICSC52841.2022.00041.
- [14] T. N. T. A. Rahim, Z. A. Aziz, R. H. A. Rauf and N. Shamsudin, "Automated exam question generator using genetic algorithm," 2017 IEEE Conference on e-Learning, e-Management and e-Services (IC3e), Miri, Malaysia, 2017, pp. 12-17, doi: 10.1109/IC3e.2017.8409231.
- [15] A. Immanuel and Tilasi.B, "Framework for Automatic Examination Paper Generation System," International Journal of Computer Science and Technology, vol. 6, no. 1, pp. 128-130, 2015.
- [16] D. L. Butler and P. H. Winne, "Feedback and Self-Regulated Learning: A Theoretical Synthesis". Review of Educational Research 65, 245–281, 1995.
- [17] S. Pandey and K. Rajeswari, "Automatic Question Generation Using Software Agents for Technical Institutions," International Journal of Advanced Computer Research, vol. 3, no. 13, pp. 307-311, 2013.