ISSN: 2710-0170

The quality evaluation index system for AI-generated digital educational resources

Yanli Chen, Chunlan Lv *

School of Economics and Management, Yibin University; Yibin, Sichuan, 644007, China

* Corresponding Author

Abstract

With the rapid development of Generative AI (GAI) technology, AI-generated content (AIGC) has become the main mode of content creation, replacing User-Generated Content (UGC) and Professional Generated Content (PGC). This transformation has led to the adoption of a brand-new form for AI-generated digital educational resources (AIGDER). However, concerns have arisen regarding the quality of generated content. To address this issue, this paper proposes a comprehensive evaluation index system for AIGDER by integrating the Delphi method and the entropy weight method. Firstly, through a systematic review of recent literature, potential quality indicators covering content, expression, user aspects, and technical aspects were identified. Subsequently, these indicators were refined through two rounds of expert consultation using the Delphi method, and finally, a structured quality indicator system was established, including four dimensions and twenty specific indicators. After that, the weight coefficients of the quality indicators were determined using the entropy weight method, and the importance of each indicator was analyzed. The proposed system provides a framework for relevant stakeholders to select high-quality AIGDER (AI educational resources) and make AI tools conform to educational standards.

Keywords

AI-generated content; digital educational resources; quality evaluation index; entropy weight method; Delphi method.

1. Introduction

In recent years, artificial intelligence (AI) has achieved significant advancements, leading to transformations across various industries and facets of human life [1]. One major area where AI has had a profound impact is content creation. With advancements and breakthroughs in generative language models, artificial-intelligence-generated content (AIGC) that supplements conventional content creation methods like user-generated content (UGC) and professional-generated content (PGC) has become a key focus worldwide [2]. AIGC refers to a novel approach to the creation of content, such as text, images, videos, and other forms of data, through artificial intelligence algorithms [3]. AIGC is achieved by extracting intent information from instructions provided by humans and generating the content according to its knowledge and the intent information [4]. In the context of AIGC, digital educational resources have entered a new paradigm of development through human-machine collaboration, which effectively addresses challenges related to quantity, quality, and efficiency in traditional resource development. Within this paradigm, the resulting resources are termed AI-Generated Digital Educational Resources (AIGDER). These refer to educational materials created using AI technologies based on human prompts or questions, capable of operating across various digital devices while supporting both teachers' instruction and students' learning processes. AIGDER primarily encompasses types such as exercises (or test questions), digital textbooks, teaching

ISSN: 2710-0170

presentations, lesson plans, and supplementary teaching materials. Looking ahead, AIGDER is poised to become a transformative form of digital educational resource development, fostering innovative multimodal learning experiences, enabling highly personalized learning pathways, and alleviating teachers' workloads. Although AIGDERs provide many benefits to education, they concurrently pose potential risks and challenges.

Cognitive biases represent a significant concern in the application of AIGDERs [5, 6]. AIGDERs utilize probabilistic models to identify patterns within textual data and automatically generate content based on syntactic rules. However, since AIGDERs fail to comprehend language semantics, they may readily generate content that is fraught with inaccuracy, bias, or unfairness [7]. A new term, “hallucination”, is used to describe the occurrence of disinformation in AIGDERs [8]. According to data from GitHub, the Chat-GPT series of platforms exhibit a 3% to 3.5% error rate, the LLaMA series of platforms have an error rate of 5.1% to 5.9%, and Google’s PaLM platform has a notably higher error rate of 12.1%. Consequently, AIGDERs can inevitably lead to cognitive biases [9]. Scholars have emphasized the necessity of cautious engagement with AIGC platforms. In addition, certain researchers have devised mathematical models to comprehensively assess both their hallucinatory tendencies and creative capabilities. However, the absence of standardized criteria and methodologies for evaluating AIGDER systems has led to inconsistent practices and operational inefficiencies. Therefore, there is an immediate need to establish a comprehensive evaluation system for AIGDERs.

To ensure the quality of AIGDERs in the context of the emerging digital world, this study proposes a framework for evaluating their quality by integrating the Delphi method with the entropy weight method. While Delphi seeks a convergence of opinions among a group of experts, the entropy weight method is mainly used to determine the weights of each indicator in a multi-index evaluation system. The integration of these two methods assists in systematically identifying the evaluation criteria and then setting priorities among them. The procedures are as follows. Through a comprehensive literature review, various dimensions and indicators consisting of content, expression, and user and technical aspects are identified. Subsequently, a two-round questionnaire is disseminated among a panel of experts to affirm the validity of each indicator via the Delphi method. Then, the entropy weight method is employed to determine the weights of the identified quality indicators, resulting in an evaluation index system that includes four dimensions and twenty indicators. The objective of this paper is to select key indicators that can inform the development of strategies for stakeholders involved in AIGDERs.

The remainder of this paper is organized as follows. Section 2 describes and details the methodology. Section 3 presents the research results. Section 4 provides the main conclusions.

2. Research Methods

The overall framework of this study is mainly divided into three steps: extraction of quality indicators, confirmation of quality indicator sets, and establishment of the quality evaluation index system, as illustrated in Figure 1.

2.1. Section Headings

To evaluate the quality of AIGDERs, the factors that affect the quality are identified using systematic literature review (SLR). SLR is a methodology that employs a standardized analytical approach to summarize the current state and development tendency within a research field or topic [10]. SLR follows a structured process proposed by Watson, which includes predefined search criteria and systematic screening processes [11]. Compared to traditional literature review, SLR has higher reliability because it addresses uncontrollable issues such as subjectivity and bias [12].

ISSN: 2710-0170

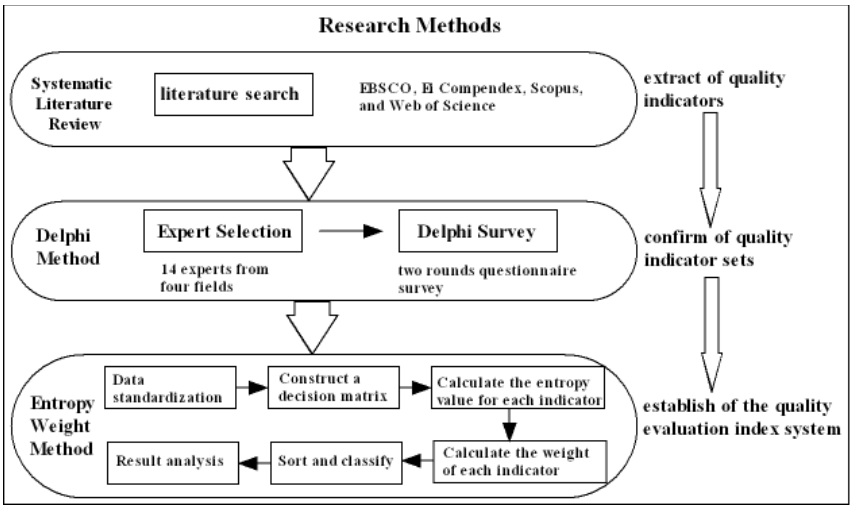

Figure 1: The overall framework of this study

According to the PRISMA guidelines, SLR typically involves searching three or more databases to retrieve literature [13]. To comprehensively acquire the necessary literature data for evaluating the quality of AIGDERs, four literature databases are selected: EBSCO, Ei Compendex, Scopus, and Web of Science. These databases extensively cover leading educational journals, and their inclusion criteria ensure that the literature within them largely meets the quality requirements of SLR. This study identified several relevant keywords from existing literature on AIGDERs and selected appropriate terms for research focus. For example, the thematic search query for the Web of Science is structured as follows: TS = (("Artificial intelligence generated content" or "AIGC" or "ChatGPT" or "Large language model*" or "Generative artificial intelligence" or "Generative AI") AND ("Education*") AND ("Resource*" OR "Material*") AND ("Quality" or "Quality evaluation" or "Evaluation index system" or "Evaluation criteria")). As ChatGPT was released by OpenAI in November 2022, the search timeframe is set from November 2022 to October 2024. For four chosen databases, appropriate filtering tools are employed to refine the research results and avoid the occurrence of duplication. Following the literature search, a total of 187 papers are retrieved from four databases after deduplication.

2.2. Delphi Method

The Delphi method, named after a Greek oracle who was famed for prophetic abilities, was initiated by the Rand Corporation in the 1950s to forecast the effect of technology on warfare [14]. The Delphi method is a systematic and qualitative method used to gather opinions from panels of selected experts to achieve consensus on outcomes and responses [15]. It is featured with anonymity and confidentiality. This method has the capability to brainstorm from experts' opinions without face-to-face contact, which avoids groupthink [16]. The Delphi method has been widely used to identify and evaluate indicators in education field. For example, Khan et al. established a consensus on the curriculum viability indicators using the modified Delphi technique to provide a framework for evaluating curriculum viability [17]. Seo et al. employed the modified Delphi method to refine key indicators for the course subject "Environment" in the secondary school curriculum in South Korea [18]. Encouraged by these studies, the Delphi method is utilized in this study to determine the indicators for the quality of AIGDERs.

2.2.1. Expert selection

According to the working procedure of the Delphi method, experts in the related fields of AIGDER were selected for investigation. Based on the opinions of the experts, the quality evaluation indicators of AIGDER were revised and improved to make them scientific and reasonable. In this study, 14 experts from four fields, namely intelligent education technology, educational digitalization, educational resource construction and development, and information technology and education application, were consulted. The research fields of the

ISSN: 2710-0170

experts were closely related to the research topic of this study, and the number of experts was within an appropriate range (8-20 people is suitable). The detailed information of the experts in each field is shown in Table 1.

Table 1: Statistics of Experts' Detailed Information

Research field	Number	Title and Education Background	Percentage
Intelligent Education Technology	2	One professor and one associate professor; both hold doctoral degrees.	14.3%
Educational Digitalization	3	Two professors and one associate professor; all hold doctoral degrees.	21.4%
Construction and Development of Educational Resources	6	There are six professors in total, among whom three are doctoral supervisors and four hold doctoral degrees.	42.9%
Information Technology and Educational Applications	3	One professor, two associate professors, and two doctoral candidates.	21.4%

2.2.2. Delphi survey

Based on the Delphi method, a questionnaire survey was conducted via email with 14 selected experts in two rounds. The first round mainly aimed at consulting the rationality of each index of resources, inviting experts to score and offer suggestions for modification. Based on the scores given by the experts in the first round and their suggestions for modification, the index system was modified accordingly. The modified index system was then subjected to the second round of expert questionnaire survey following the procedure of the first round. During this process, the authority coefficient (Cr) of the experts was used to measure their familiarity with the field and the credibility of their questionnaire survey results. It was mainly calculated by taking the arithmetic average of the experts' familiarity with the indicators (Ca) and the judgment basis (Cs) [14].

2.3. Entropy Weight Method

The Entropy Weight Method (EWM) is a method used to determine the weights of various indicators in a multi-index evaluation system [19]. In multi-attribute decision-making analysis, different indicators have varying degrees of influence on the overall evaluation result. Therefore, weights need to be set. The core idea of the Entropy Weight Method is to measure the information content of each indicator based on the size of information entropy, and thereby determine the importance of the indicators. The smaller the entropy value, the greater the information content, and the higher the weight should be. This method can effectively avoid the interference of subjective factors and make the determination of weights more scientific and objective.

2.3.1. Information entropy

The concept of entropy originates from physics and information theory. In information theory, entropy (Entropy) is used to measure the degree of uncertainty of a system. In decision analysis, if the data variance of an indicator is large, it implies that it provides a great deal of information; conversely, if the data variance is small, it indicates that it provides very little information [19].

2.3.2. Calculation steps

The basic calculation steps of the entropy weight method are as follows [19]:

Step 1: Data Standardization

Due to the possible differences in the units and ranges of values of various indicators, in order to eliminate such discrepancies, data standardization processing is required. Common methods

ISSN: 2710-0170

include linear dimensionless methods, which normalize the indicator values within the range of [0, 1].

Standardization formula for positive indicators:

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \tag{1}$$

Standardization formula for negative indicator:

$$x'_{ij} = \frac{\max(x_j) - x_{ij}}{\max(x_j) - \min(x_j)} \tag{2}$$

Step 2: Calculate the information entropy

First, calculate the proportion P_{ij} of the index value of the $i-th$ scheme under the $j-th$ indicator:

$$P_{ij} = \frac{x'_{ij}}{\sum x'_{ij}} \tag{3}$$

Then calculate the entropy value E_j for the $j-th$ item indicator:

$$E_j = -\frac{1}{\ln(n)} \sum_i P_{ij} \ln(P_{ij}) \tag{4}$$

Here, n represents the sample size, j is a fixed value, and $\ln$ stands for natural logarithm.

Step 3: Determine the weights

First, calculate the variance index D_j for the $j-th$ indicator:

$$D_j = 1 - E_j \tag{5}$$

The larger the variation index is, the greater the difference of the evaluated object it indicates, and thus the greater the contribution of this index to the evaluation result.

Then calculate the weight W_j of the $j-th$ item indicator:

$$W_j = \frac{D_j}{\sum D_j} \tag{6}$$

Step 4: Calculate the comprehensive score

The comprehensive evaluation value Z_i of the $i-th$ evaluation object is:

$$Z_i = \sum W_j * P_{ij} \tag{7}$$

By following the above steps, we can calculate the comprehensive scores of each evaluation object using the entropy method, and thereby rank or classify them in terms of their merits and demerits.

3. Results

3.1. The Initial Identified Quality Indicators

Using the SLR, Table 2 illustrates the four dimensions for coding and offers a description of the quality indicators coded within each dimension. Information is collected across the four dimensions for coding: content characteristics, expression characteristics, user characteristics, and technical characteristics. Specifically, content characteristics are derived from a comprehensive review of the existing literature on high-quality educational content, such as [20]. Expression characteristics are identified through research on the expression of AIGC in educational settings, such as [21]. User characteristics are developed based on user-centered design principles and research on how users interact with AIGC, such as [22]. Technical

ISSN: 2710-0170

characteristics are determined by considering the technological requirements of modern educational tools, such as [23]. For each indicator in the four dimensions, the frequency of its occurrence in the literature is counted. The codes are ranked based on their frequency, providing a clear picture of the relative importance of each theme. These codes are then used as input indicators for the Delphi and Entropy weight methods in the subsequent methodological steps.

Table 2: List of codes.

Dimension	Specific Coding Content
Content characteristics	Authenticity, accuracy, specifications, relevance, novelty, diversity, timeliness
Expression characteristics	Legitimacy, knowledgeability, logicality, comprehensible
User characteristics	Achievement, acquisition, compatibility, friendliness
Technical characteristics	Conciseness, stability, human-analogy, security, big data

3.2. Analysis of the Delphi Survey

Based on the Delphi method, a questionnaire survey was conducted via email with 14 selected experts in two rounds. The first round mainly aimed at consulting the rationality of each indicator, inviting experts to score and offer suggestions for modification. Then, based on the scores given by the experts in the first round and their suggestions for modification, the indicator system was modified accordingly. Subsequently, the modified indicator system was subjected to the second round of expert questionnaire survey following the same procedure as the first round. Through calculation, it can be concluded that the Cr values of the first and second rounds of this study were 0.726 and 0.789 respectively (both greater than 0.7). Therefore, the results of this expert questionnaire survey have high authority and credibility.

Through two rounds of expert inquiries, the evaluation index system was revised based on the scores given by experts on the rationality of the indicators and the suggestions they put forward for modification. The “diversity” indicator is excluded due to its conceptual overlap with other metrics and insufficient specificity for targeted evaluation. Similarly, the “human-analogy” indicator is eliminated as it may not guarantee the fairness of the quality judgment. To enhance terminological precision, the “knowledgeability” indicator is revised to the “normativity” indicator to emphasize adherence to established standards, while the “specifications” indicator is redefined as the “disciplinarity” indicator to clarify alignment with domain-specific rigor. The “computation power” indicator is introduced to evaluate generative AI’s functional efficiency in processing and synthesizing data. The “experience feeling” indicator is incorporated to underscore the pedagogical significance of evaluating learning processes alongside outcomes. Ultimately, four first-level indicators and twenty second-level indicators were obtained, forming the AIGDER quality evaluation index system (as shown in Figure 2).

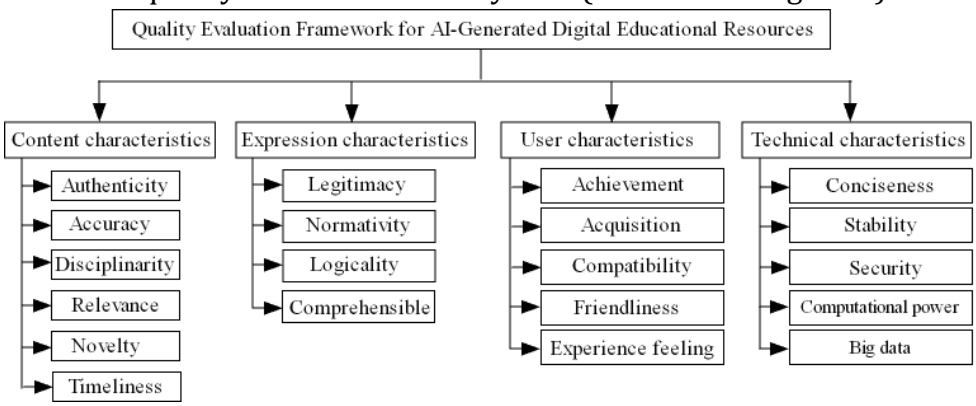

Figure 2: The AIGDER quality evaluation index system

ISSN: 2710-0170

3.3. Determination of the Weight for Identified Quality Factors Using EWM

The EWM is used to determine the weights for the quality of AIGDERs. Take an example of an expert's consultation as follows. The evaluation scores (0-100) of the four primary indicators and the twenty secondary indicators by him are shown in Table 3.

Table 3: Evaluation Index Score Table for 20 Secondary Indicators

First-Level Indicator	Second-Level Indicator	Score
Content characteristics	Authenticity	84
	Accuracy	88
	Disciplinarity	87
	Relevance	91
	Novelty	80
	Timeliness	90
Expression characteristics	Legitimacy	85
	Normativity	90
	Logicality	87
	Comprehensible	78
User characteristics	Achievement	86
	Acquisition	83
	Compatibility	83
	Friendliness	83
	Experience feeling	87
Technical characteristics	Conciseness	85
	Stability	90
	Security	79
	Computational power	91
	Big data	90

Based on the original scoring table, the data are standardized first, then the information entropy of each indicator is calculated, and finally the weights of each indicator are determined. As shown in Table 4.

Table 4: The quality evaluation index system of AIGDERs.

First-Level Indicator	First-Level Weight	Second-Level Indicator	Second-Level Weight	Rank
Content characteristics	0.2639	Authenticity	0.0973	1
		Accuracy	0.0360	16
		Disciplinarity	0.0320	19
		Relevance	0.0284	20
		Novelty	0.0575	7
		Timeliness	0.0459	9
Expression characteristics	0.2370	Legitimacy	0.0642	5
		Normativity	0.0425	11
		Logicality	0.0321	18
		Comprehensible	0.0652	4

ISSN: 2710-0170

User characteristics	0.2120	Achievement	0.0390	14
		Acquisition	0.0462	8
		Compatibility	0.0420	12
		Friendliness	0.0405	13
		Experience feeling	0.0443	10
Technical characteristics	0.2870	Conciseness	0.0693	3
		Stability	0.0816	2
		Security	0.0369	15
		Computational power	0.0350	17
		Big data	0.0641	6

4. Conclusion

The increasing incorporation of AIGDERs within university settings has become a notable trend. Concurrently, this development has sparked concerns regarding the quality of AIGDERs. To ensure the use of high-quality AIGDER, this paper has developed an evaluation index system for AIGDER by combining the Delphi method and the EWM method, which consists of four dimensions and twenty indicators. The findings underscored the paramount importance of technical characteristics in the quality assessment of AIGDERs. Moreover, content characteristics were considered to be the second most important factor, and the importance of expression and user characteristics was properly recognized. Among the second-level indicators, the experts regarded “Authenticity”, “Stability”, “Conciseness”, and “Comprehensible” more important than other indicators. The proposed system provided relevant stakeholders with a framework to select high-quality AIGDERs and guide AI tools towards educational standards.

References

[1] Wang, W. and K. Siau, Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work and Future of Humanity: A Review and Research Agenda. Journal of Database Management, 2019. 30(1): p. 61-79.

[2] Lin, H., et al., Comparing AIGC and traditional idea generation methods: Evaluating their impact on creativity in the product design ideation phase. Thinking Skills and Creativity, 2024. 54(000).

[3] Xu, M., et al., Unleashing the Power of Edge-Cloud Generative AI in Mobile Networks: A Survey of AIGC Services. Ieee Communications Surveys and Tutorials, 2024. 26(2): p. 1127-1170.

[4] Cao, Y., et al., A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT arXiv. arXiv, 2023.

[5] Lambert, J. and M. Stevens, ChatGPT and Generative AI Technology: A Mixed Bag of Concerns and New Opportunities. Computers in the Schools, 2024. 41(4): p. 559-583.

[6] Wu, T.-T., et al., Promoting Self-Regulation Progress and Knowledge Construction in Blended Learning via ChatGPT-Based Learning Aid. Journal of Educational Computing Research, 2024. 61(8): p. 3-31.

[7] Kortemeyer, G., Could an artificial-intelligence agent pass an introductory physics course? Physical Review Physics Education Research, 2023. 19(1).

[8] Sun, Y., et al., AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanities & Social Sciences Communications, 2024. 11(1).

ISSN: 2710-0170

[9] Watts, F.M., et al., Comparing Student and Generative Artificial Intelligence Chatbot Responses to Organic Chemistry Writing-to-Learn Assignments. Journal of Chemical Education, 2023. 100(10): p. 3806-3817.

[10] Jing, Y., et al., Bibliometric mapping techniques in educational technology research: A systematic literature review. Education and Information Technologies, 2024. 29(8): p. 9283-9311.

[11] Labadze, L., M. Grigolia, and L. Machaidze, Role of AI chatbots in education: systematic literature review. International Journal of Educational Technology in Higher Education, 2023. 20(1).

[12] Waqar, A., et al., Applications of AI in oil and gas projects towards sustainable development: a systematic literature review. Artificial Intelligence Review, 2023. 56(11): p. 12771-12798.

[13] Moher, D., et al., Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Bmj-British Medical Journal, 2009. 339.

[14] Dalkey, N. and O. Helmer, AN EXPERIMENTAL APPLICATION OF THE DELPHI METHOD TO THE USE OF EXPERTS. Management Science, 1963. 9(3): p. 458-467.

[15] Olsen, A.A., et al., How to use the Delphi method to aid in decision making and build consensus in pharmacy education. Currents in Pharmacy Teaching and Learning, 2021. 13(10): p. 1376-1385.

[16] Zhao, P., Z.M. Ali, and Y. Ahmad, Developing indicators for sustainable urban regeneration in historic urban areas: Delphi method and Analytic Hierarchy Process (AHP). Sustainable Cities and Society, 2023. 99.

[17] Khan, R.A., et al., Curriculum Viability Indicators: A Delphi Study to Determine Standards and Inhibitors of a Curriculum. Evaluation & the Health Professions, 2021. 44(3): p. 210-219.

[18] Seo, E., J. Ryu, and S. Hwang, Building key competencies into an environmental education curriculum using a modified Delphi approach in South Korea. Environmental Education Research, 2020. 26(6): p. 890-914.

[19] Guo, Y. and J. Li, Application of tender evaluation model based on entropy weight to electric power construction projects. Engineering Journal of Wuhan University, 2013. 46(2): p. 184-7.

[20] Tanaka, O.M., et al., Assessing the reliability of ChatGPT: a content analysis of self-generated and self-answered questions on clear aligners, TADs and digital imaging. Dental press journal of orthodontics, 2023. 28(5): p. e2323183-e2323183.

[21] Liu, X. and Y. Xiao, Chinese university teachers' engagement with generative AI in different stages of foreign language teaching: A qualitative enquiry through the prism of ADDIE. Education and Information Technologies, 2025. 30(1): p. 485-508.

[22] Maheshwari, G., Factors influencing students' intention to adopt and use ChatGPT in higher education: A study in the Vietnamese context. Education and Information Technologies, 2024. 29(10): p. 12167-12195.

[23] Lu, G., N.B. Hussin, and A. Sarkar, Navigating the Future: Harnessing Artificial Intelligence Generated Content (AIGC)for Enhanced Learning Experiences in Higher Education. 2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science. 2024. 1-12.