Hybrid Genetic Adaptive Test Paper Generation Algorithm with Deep Learning for Online Testing Systems

Jiaquan Zhu, Haiqing Wei*
Department of Commerce and Management, Guangxi Natural Resources Vocational and Technical College Chongzuo 532199, China
E-mail: Whqzjq875204@163.com, 15278789967@163.com
*Corresponding author

Keywords: deep learning, automatic test paper generation algorithm, line testing system, genetic algorithm, adaptability

Received: October 25, 2024

Socio-economic development has brought about the expansion of education scale. Due to the tasks of printing test papers, proctoring, and grading, teachers' workload is increasing, and these processes are all manually completed, which is very prone to errors. Consequently, this paper proposes an online test system that integrates deep learning and an automatic test composition algorithm. The objective of this integration is to automatically generate test papers that align with teaching requirements through intelligent means. This approach aims to minimize manual intervention, thereby reducing error rates and enhancing test efficiency. Firstly, the paper studies the automatic paper composition algorithm based on deep learning and optimizes the paper composition efficiency by intelligently reducing the search space, setting constraints, and using real number segmentation coding. Then, an online test system is designed to automatically record answers and score them. Finally, the results of the online test system combining deep learning and automatic test composition algorithm are analyzed. The results showed that when the evolutionary algebra reached about 400, the population maximum fitness value of the studied algorithm was stable at 18. In 20 independent experiments, the algorithm showed excellent convergence performance. The convergence algebraic curve of the first experiment tended to be stable in about 180 generations, and the shortest running time of the algorithm was only 0.5 seconds. The average accuracy of the research algorithm was 92% in the difficult test task, and 93% in the test task, which fully verified the efficiency and stability of the algorithm. With the increase in population size, the mean fitness of the online test system proposed by the study also increased. When the individual population size was 300, the mean fitness of the online test system proposed by the study was 17.173, the mean fitness of the exam treasure was 17.162, and the mean fitness of the exam star was 17.158. The mean fitness of learning Xiaoyi was 17.153. The superior optimization ability and efficient convergence performance of the online test system can provide strong support for the rational allocation of educational resources and the realization of personalized teaching.

Povzetek: Predstavljeni hibridni algoritem združuje globoko učenje in genetske pristope za samodejno sestavljanje testov. Z uporabo transformerjev skrči iskalni prostor, segmentno kodiranje varuje omejitve, hibridni operaterji pospešijo konvergenco.

1 Introduction

The rapid development of science and technology and the continuous deepening of educational reform have posed many challenges to traditional teaching and examination models. The conventional approach to test composition involves the manual screening of test questions, which is susceptible to subjective influences. This method often leads to variations in the difficulty levels of test questions, incomplete coverage of knowledge points, and the repetition of questions [1-2]. In the contemporary context of constrained educational resources and heterogeneous student requirements, the efficient and precise generation of test materials that align with pedagogical imperatives has emerged as a pressing concern that demands immediate attention within the domain of educational technology. Deep Learning (DL), as a branch of artificial intelligence, has made breakthrough progress in various fields like image recognition and natural language processing in recent years [3-4]. For example, Lundine et al. applied DL to underwater fluid problems, which can be used for rapid detection and characterization in situations where other high-resolution depth measurement methods are available. The value of data-driven detection models in characterizing the complex morphology of the seabed has been demonstrated [5]. Bitachon et al. proposed a method to reduce the training time of DL-based digital backpropagation and applied it to channel wavelength

division multiplexing transmission links. This scheme could improve the compensation performance to 0.48dB by training only on the last segment of the stacked DL-DBP [6]. In the domain of data mining, Wang et al. have proposed a network intrusion detection model based on DL, which has been shown to enhance the accuracy and efficiency of intrusion detection by optimizing feature representation and combining Convolutional Neural Networks (CNNs) and transformer architectures [7]. Latreche et al. reviewed EEG signal analysis based on DL. The DL model could extract features related to fatigue state from EEG signals and achieve high-precision fatigue detection [8]. The Test Paper Generation (TPG) algorithm randomly selected test questions from the question bank until the test paper was completed or it was no longer possible to continue extracting test questions that met the conditions. This method had a simple structure and fast speed. Therefore, Yuan adopted a random TPG algorithm for automatic grading of essays, which can accurately score students' essay grades [9]. Therefore, combining DL with an automatic TPG algorithm, this study proposes a Hybrid Genetic Adaptive TPG Algorithm (HGATPGA) aimed at in-depth mining and analysis of test data, achieving automated generation and intelligent management of test papers.

While DL has made breakthroughs in several fields, existing methods still have limitations. For example, the application of traditional DL models on large-scale data sets is limited by high memory and computational requirements. In addition, many existing methods perform poorly in optimizing efficiency and scalability, especially in dynamic data environments and large-scale distributed systems. To more intuitively show the gap between the existing methods and the most advanced methods, the performance of HGATPGA is compared with that of Niched Genetic Algorithms (NGA) and Adaptive Genetic Algorithms (AGA). The comparison results are shown in Table 1.In Table 1, HGATPGA is superior to NGA and AGA in terms of accuracy, running time, and adaptability. HGATPGA significantly improves its scalability on large-scale datasets through efficient feature extraction, adaptive optimization strategy, efficient feature dimension reduction, and distributed computing support of DL models. When dealing with large-scale datasets, NGA and AGA have obvious shortcomings in optimization efficiency and scalability. In the context of high-dimensional data, the computational complexity of NGA escalates considerably, leading to extended execution times. When AGA is extended to larger datasets, it shows insufficient scalability, which makes it difficult to adapt to the needs of the modern big data environment.

Table 1: Summary of advanced method performance

Method	Accuracy (%)	Run time (s)	Fitness value	Application scenario
HGATPGA	98.5	120	0.92	Complex optimization problem
NGA	93.7	150	0.89	Large-scale data processing
AGA	94.5	130	0.91	Real-time system optimization

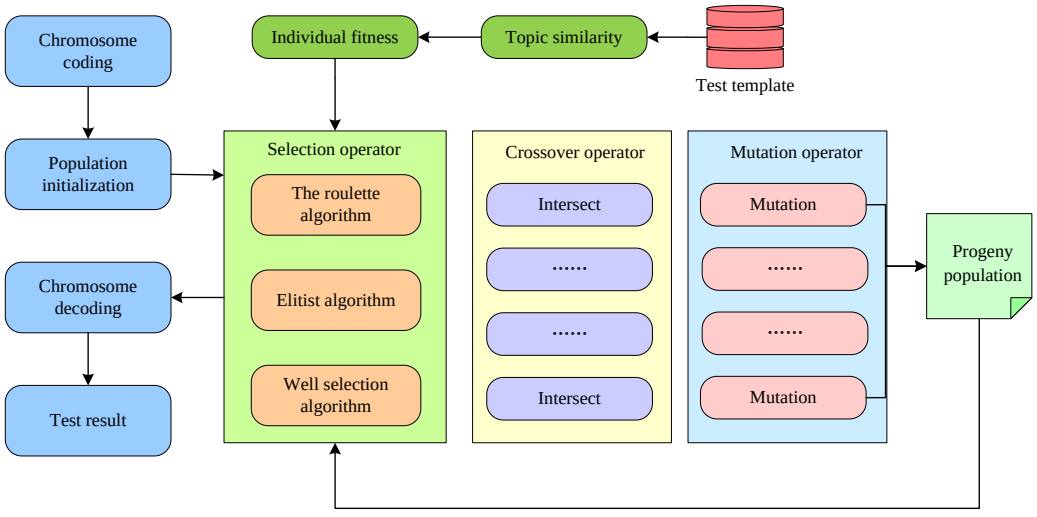


Figure 1: Automatic TPG system overall design flow

The innovation of this study lies in the use of DL models for deep mining of massive test data, automatically extracting key features such as test difficulty, knowledge points, and question types. This study uses a Transformer pre-trained language model to encode test text and transform it into a high-dimensional vector representation. Furthermore, the Transformer architecture's parallel processing capability enhances the efficiency of feature extraction, thereby enabling the system to process large-scale item data with optimal efficiency. The implementation of this research not only helps to reduce the workload of teachers and improve the

efficiency of teaching resource utilization but also promotes innovation and development of educational technology, providing strong support for the intelligent transformation of the education field. It is hoped that this research will result in more efficient, convenient, and personalized learning and examination experiences for students and teachers.

2 Methods and materials

2.1 Automatic TPG algorithm integrating DL

To improve the constraint handling of DL and Genetic Algorithm (GA) generated papers, this paper studies the use of DL models to intelligently reduce the search space, thereby reducing unnecessary calculations and setting constraints. The automatic TPG function refers to the online examination system or TPG software that automatically randomly or intelligently selects questions from the question bank based on preset conditions such as question bank, question type, quantity, difficulty, and knowledge points to form a complete test paper [10]. The design process of automatic TPG system is displayed in Figure 1.

In online testing, the effectiveness of the test is usually scientifically measured and evaluated using four indicators: difficulty, validity, reliability, and discrimination. The test difficulty is mainly segmented into the difficulty of objective questions, the difficulty of subjective questions, and the overall difficulty of the test paper [11]. Among them, the difficulty of objective questions refers to the ratio of the number of people who answer the question correctly to the total number of people participating in the test. The expression for objective question difficulty DIF_o is shown in equation (1).

$$DIF_o = 1 - \frac{R}{N} \tag{1}$$

In equation (1), R is the number of people who answered the question correctly. N is the overall people who answered this question. Subjective difficulty usually refers to the difficulty level of a test, and for ability tests, the difficulty level of an item can be measured by its difficulty [12]. The expression for subjective difficulty DIF_s is shown in equation (2).

$$DIF_s = 1 - \frac{\bar{P}}{M} \tag{2}$$

In equation (2), $\bar{P}$ is the average score of the question. M is the total score of this question. The test content consists of objective and subjective question types. The overall difficulty of the test content should take into account the difficulty of the test questions. The expression for the overall difficulty DIF_p is shown in equation (3).

$$DIF_p = \frac{\sum_{i=1}^{n} d_i F_i}{\sum_{i=1}^{n} F_i} \tag{3}$$

In equation (3), i is the test question number. n is the total number of questions in the exam paper. d_i and F_i are the difficulty and score of each question. Test validity refers to the effectiveness of a test, which means the degree to which a set of tests corresponds to the content being tested [13]. The calculation formula for validity V is shown in equation (4).

$$V = \frac{\sum_{e=1}^{n} (a_e - \bar{a})(b_e - \bar{b})}{lS_a S_b} \tag{4}$$

In equation (4), l is the total number of participants in the test. a and b are the average scores of all candidates in two exams. $S_a S_b$ represents the standard deviation of all test takers' scores on both tests. Test discrimination is the ability of a test question to distinguish the situation of a subject, mainly used to evaluate topics selected for selection purposes [14]. The formula is shown in equation (5).

$$d = \frac{h - l}{k} \tag{5}$$

In equation (5), h and l are the high-score and low-score groups. k is the actual score of the test. When constructing the test paper, each question is determined by the above five indicators, forming a 5-dimensional vector matrix expression as shown in equation (6).

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{15} \\ a_{21} & a_{22} & \cdots & a_{25} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{m5} \end{bmatrix} \tag{6}$$

In equation (6), a_{m1} is the test question number. a_{m2} is the test score. a_{m3} is the question type. a_{m4} is the chapter. a_{m5} is the difficulty. In automatic TPG systems, traditional methods such as random sampling and backtesting have their limitations. This study uses GA for generating test papers, but its convergence speed is slow and the computational cost is high. Therefore, this study utilizes DL models to intelligently narrow down the search space, reduce unnecessary computational complexity, and set constraints. To make sure the success of the test paper, it is needed to conform to five constraints: test question number, score, question type, chapter, and difficulty [15]. Among them, the constraint formula for the distribution of the total Score of the Test Paper (TPS) is shown in equation (7).

$$M = \sum_{m=1}^{n} a_{m2} \tag{7}$$

The constraint equation for the distribution of question types is shown in equation (8).

$$T_t = \sum_{m=1}^{n} B_3 a_{m3} \tag{8}$$

The constraint formula for chapter score distribution is shown in equation (9).

$$Z_t = \sum_{m=1}^{n} B_4 a_{m4} \tag{9}$$

The constraint formula for difficulty score distribution is

shown in equation (10).

$$N_t = \sum_{m=1}^{n} B_5 a_{m5} \tag{10}$$

GA usually uses binary encoding, and using binary variations during TPG can greatly reduce encoding efficiency [16]. This study adopts real number segmented encoding, with the real numbers of the test questions as genes, and segmented encoding according to the types of test questions. The expression for the encoding scheme is shown in equation (11).

$$P = \left\{ Q1_1 ... Q1_{n1} \middle| Q2_1 ... Q2_{n2} \middle| ... \middle| Q1_1 ... Qt_{nt} \right\} \tag{11}$$

In equation (11), Q_i is the i-th type of question in the test paper. t is the total number of question types. n_i is the quantity of each type of question. In the process of constructing the mathematical model of paper generation, it is essential to ensure that the generated paper can accurately meet the teaching requirements and have good quality. In this study, five constraint conditions, a_{m1}, a_{m2}, a_{m3}, a_{m4}, and a_{m5} are selected when establishing the mathematical model. Based on the expression of the objective function of the established mathematical model, equation (12) is obtained.

$$\min f(x) = \omega_1 m_1 + \omega_2 m_2 + ... + \omega_n m_n \tag{12}$$

In equation (12), m_n is the average value of the error between the actual score extracted from the set constraint conditions and the previously set expected score [17]. To improve the efficiency of GA in automatic test composition system, the design of fitness function is very important. The intricacy of fitness functions exerts a direct influence on the calculation time of GA, necessitating the maximization of simplification to mitigate this complexity. The fitness function should directly reflect the constraints and optimization objectives of the test paper, such as question type distribution, difficulty control, and knowledge point coverage [18-19]. The fitness function is shown in equation (13).

$$f = \sum_{i=1}^{5} \omega_i f_i + c \tag{13}$$

c is used to ensure the non negativity of the objective function. ω_i is the weight of each constraint condition in the objective function. The selection probability is shown in equation (14).

$$W_i = \frac{f(c_i)}{\sum_{j=1}^{N} f(c_i)} \tag{14}$$

In equation (14), c_i is the test question, $\frac{f(c_i)}{\sum_{j=1}^{N} f(c_i)}$ is the fitness value of c_i, and $\sum_{j=1}^{N} f(c_i)$ is the fitness value of all test questions. The calculation formula for c_i is shown in equation (15).

$$c_i = \frac{g(x_i)}{\sum_{i=1}^{r} g(x_i)} \tag{15}$$

In equation (15), $g(x_i)$ is the fitness value of the i-th individual. r is the size of the population. To maintain the constraints on the type and number of small questions in each item segment, an innovative chromosome crossover strategy is proposed. This strategy can significantly improve the adaptability and efficiency of the algorithm by introducing a new crossover mechanism. The chromosome crossover strategy is shown in Figure 2. In Figure 2, this study integrates single-point crossing and multi-point crossing methodologies to execute single-point crossing operations at the boundaries of two distinct types of question segmentation. This approach is employed to guarantee that the continuity and integrity of the question structure remain uncompromised. Within each independent question type segment, multi-point crossover techniques are used to promote diverse combinations and optimization adjustments of sub-questions within that segment. While simplifying the operation process, this study adopts an intuitive mutation operator design, randomly selecting small questions as mutation focuses to simulate gene mutation phenomena in the process of biological evolution. The process of generating test papers entails the identification of the maximum value of the fitness function. The evolution of the population ceases once the optimal individual within the population fulfills the established requirements or attains a specified number of evolutions. The optimal individual is the test paper that meets the predetermined goals. The automatic TPG flowchart based on GA is shown in Figure 3.

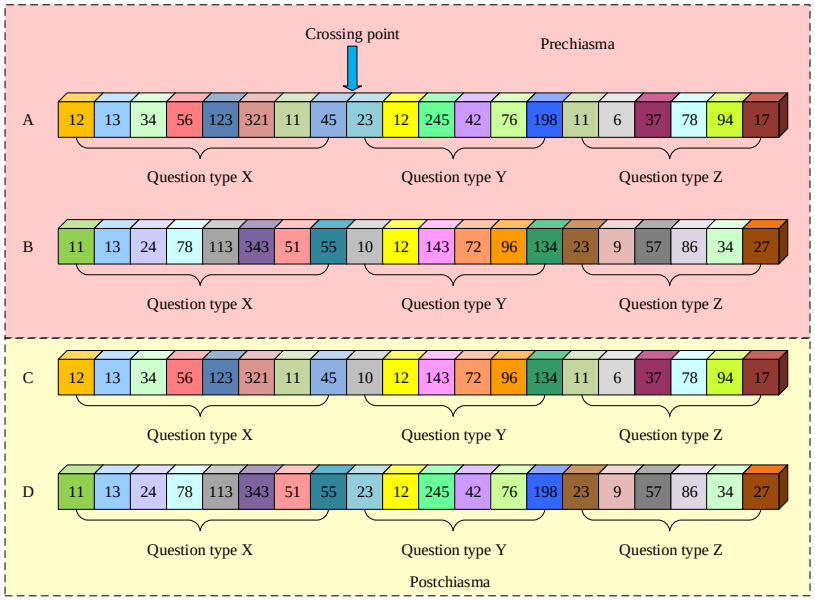


Figure 2: Cross operator diagram

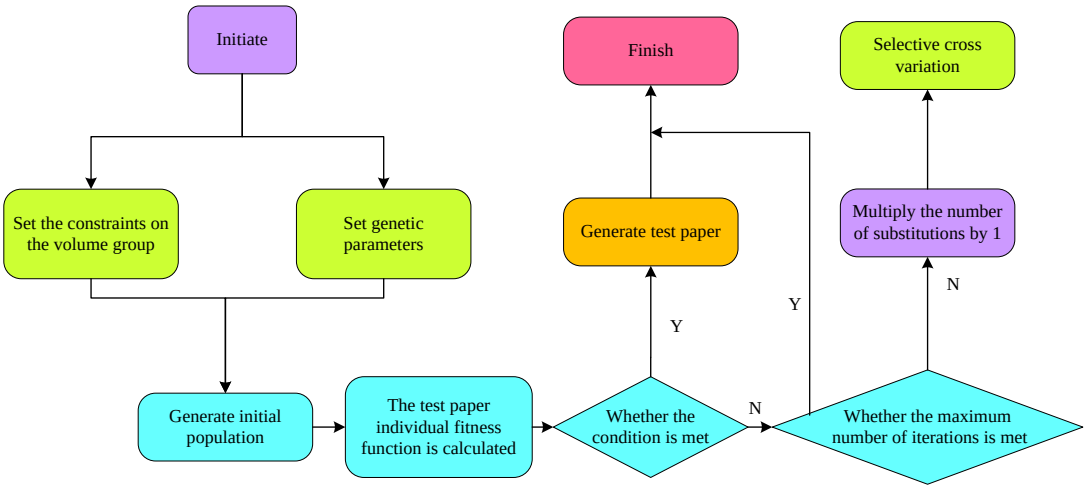


Figure 3: Flow chart of adaptive TPG algorithm based on hybrid GA

In Figure 3, constraint conditions and parameters of GA are first set for the system, and then question samples are randomly selected from the test question bank to build the initial test paper population. The system uses the fitness evaluation mechanism to evaluate each initial test paper. If the fitness value of a certain test paper directly meets the preset threshold, it is immediately regarded as a qualified test paper and output; if it does not meet the preset threshold, it enters the genetic optimization process. In the genetic optimization stage, the research carries out cross operations according to cross probability, implements multi-point crosses inside each question type, and then randomly adjusts the position of the questions in the test paper according to the mutation probability. This process is intended to increase the diversity of the population and explore new solution space. The initial size of the population is determined according to the size of the test bank and the complexity of the test paper, and is set to 10%-20% of the number of questions to ensure the diversity of the population and search efficiency. The evolution process halts once the fitness value of the optimal individual in the population attains the preset threshold or the specified number of generations 1000. The optimal individual is the test paper that fulfills the preset objective. The pseudo-code for HGATPGA is shown in Figure 4.

```
1. Initialize:
   1.1. Set population size (POP_SIZE), maximum generations (MAX_GEN), crossover rate (CROSS_RATE), mutation rate (MUT_RATE), tabu list size (TABU_SIZE), and other parameters.
   1.2. Generate an initial population of solutions (POP) randomly or using a heuristic.
   1.3. Initialize the tabu list (TABU_LIST) as empty.
   1.4. Evaluate the fitness of each solution in POP.

2. For generation = 1 to MAX_GEN:
   2.1. Selection:
      2.1.1. Select parents from POP using a selection method (e.g., tournament selection or roulette wheel).
   2.2. Crossover:
      2.2.1. Perform crossover on selected parents with probability CROSS_RATE to generate offspring.
   2.3. Mutation:
      2.3.1. Perform mutation on offspring with probability MUT_RATE.
   2.4. Evaluate the fitness of offspring.
   2.5. Tabu Search:
      2.5.1. For each offspring:
         a. Generate neighboring solutions.
         b. Evaluate the fitness of neighboring solutions.
         c. Select the best non-tabu neighbor (not in TABU_LIST).
         d. Update the tabu list (TABU_LIST) with the selected move.
         e. Replace the offspring with the best neighbor if it improves fitness.
   2.6. Replacement:
      2.6.1. Replace the worst solutions in POP with the improved offspring.
   2.7. Update the best solution found so far.
3. Output:
   3.1. Return the best solution found.
```

Figure 4: HGATPGA pseudo-code

2.2 Implementation of OTS based on HGATPGA

The Online Testing System (OTS) is mainly implemented by two subsystems, namely the TPG system and the OTS. The main function of the TPG system is to generate test papers. It randomly selects questions based on user needs such as difficulty, quantity, units, and other parameters to form a test paper. The main job of the online testing subsystem is to automatically record and score the answers after the test is completed. Specifically, the sub-system can collect the response data of the candidates in real-time, analyze and evaluate the answer results, and quickly generate a score report according to the preset scoring criteria. The structure of the OTS based on HGATPGA is shown in Figure 5.

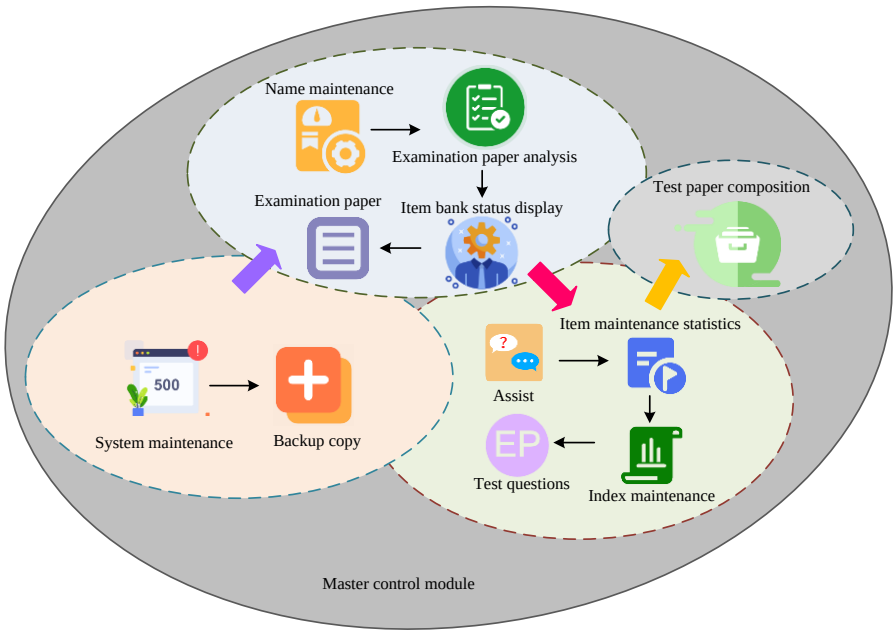


Figure 5: Structure diagram of online test system based on HGATPGA

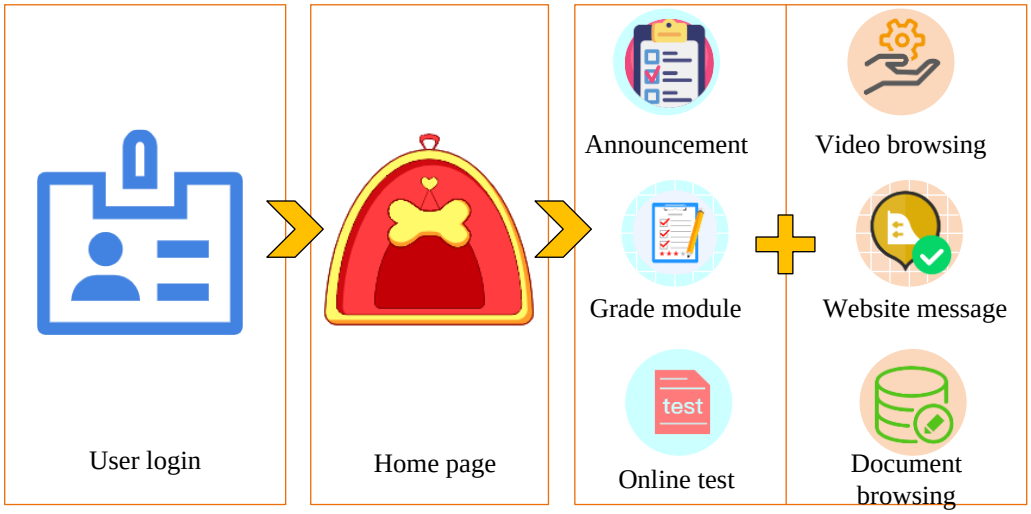


Figure 6: Tester functional module design diagram

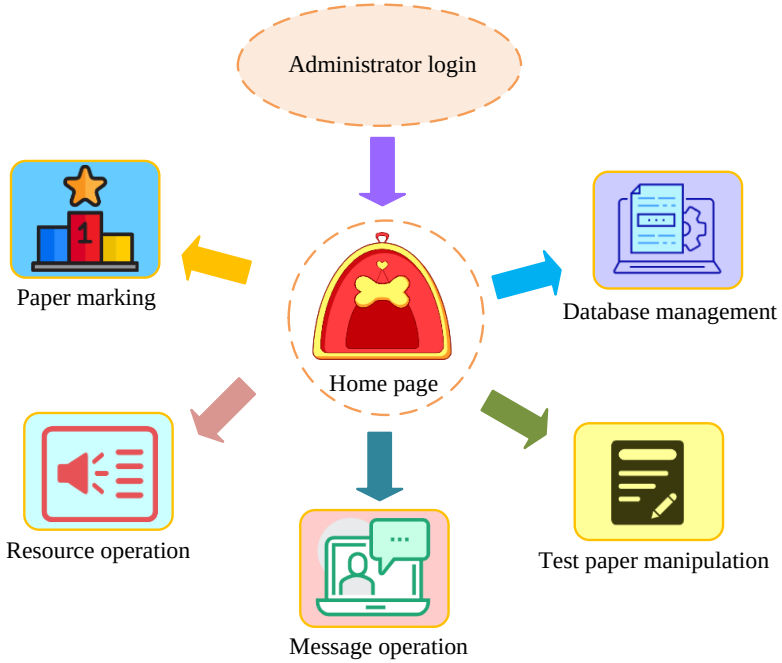


Figure 7: Administrator function module design schematic diagram

In Figure 5, in the data maintenance module, data entry is done through file or text input to meet the different usage habits and actual needs of users. The data modification process involves several steps, including the verification of consistency in restrictions pertaining to the addition and modification of information. It also encompasses the testing of the accuracy of data displayed on the modified page and the synchronization of updates to associated modules following the successful modification of data. The data are deleted through processes such as confirmation of deletion prompts, deletion constraint checks, data integrity checks, permission checks, deletion methods, data visibility, special case testing, and interface update checks. The data query module allows users to comprehensively view the entered dataset. The system maintenance module includes five sub-modules: data backup, data import, code maintenance, password change, and system initialization. The data backup function ensures data security. The data introduction module enables fast recovery of data. The code maintenance module allows for adding, deleting, and modifying code elements in the system. The password change function protects user privacy by setting parameters and entering basic files and initial data based on the actual situation of the test content. The database module of the TPG system includes a course library, a teaching outline library, and a test question indicator library. Figure 6 is a design diagram of the tester subsystem.

In Figure 6, the administrator posts various notifications in the announcement module, allowing users to promptly access and respond to learning updates. In the document browsing module, users can freely access and download diverse documents stored on the server. In the video

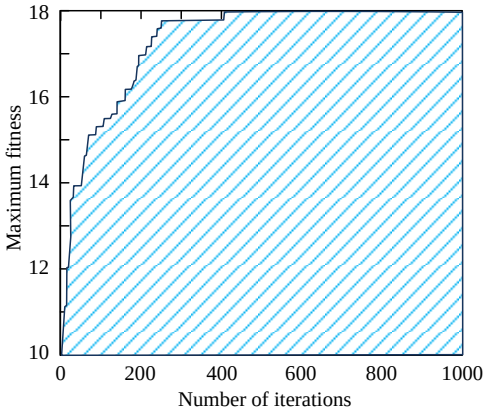
browsing module, users can watch or download teacher lecture videos online to enhance learning effectiveness. The online testing module supports independent simulation testing and supervised formal exams. The score module can help users review their learning process, identify knowledge blind spots, and provide data support for targeted learning in the future through score recording and error statistics. The website message module is used to encourage users to leave comments and suggestions on documents, video resources, and message board content to enhance the activity of the learning community. The administrator sub-system includes tasks for uploading and maintaining various learning resources such as documents and videos, and is fully responsible for database management and optimization. Figure 7 is a design diagram of the administrator function module. In Figure 7, in the paper scoring module, the administrator scores complex question types such as essay questions in the online test paper based on the scoring criteria. The resource operation module enables administrators to undertake actions such as adding, deleting, and modifying key learning resources, including news announcements, video tutorials, and document materials. This ensures the timeliness and richness of the resource library, thereby meeting the diverse learning needs of users. In the message operation module, administrators can promptly review and process user feedback, make necessary comments, modifications, or deletions to maintain a good interactive atmosphere, promote continuous optimization of platform functions, and improve user experience. In the test paper operation module, users can generate test papers in accordance with their requirements. For automatically generated test papers, administrators have modification permissions to ensure that the test paper content matches the teaching objectives. In the database management module, administrators perform maintenance operations on the system to ensure the accuracy of database information and lay a solid foundation for the stable operation of the entire online learning platform.

3 Results

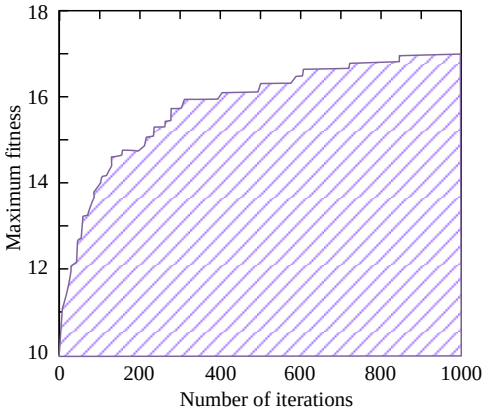
3.1 HGATPGA performance testing

Under the same conditions, this study conducts comparative experiments between NGA and AGA as comparison algorithms and HGATPGA. NGA is an improved GA based on niche technology. By introducing a niche mechanism, NGA can maintain population diversity and avoid premature convergence, so that multiple optimal solutions or near-optimal solutions can be found simultaneously. The improved GA is characterized by its capacity for dynamic adjustment of genetic parameters, contingent on the evolutionary state of the population. This capability enhances the algorithm's global search capability and convergence speed. A test question dataset containing 10,000 questions is used in the experiment. The data types are single-choice, multiple-choice, fill-in-the-blank, and short-answer. The number of each question type is 4,000, 3,000, 2,000, and 1,000 respectively. The difficulty level of the questions is divided into three categories: low, medium, and high. The low-difficulty questions account for 40% of the total, the medium-difficulty questions account for 40%, and the high-difficulty questions account for 20%. By comparing the population maximum fitness curves of different algorithms, this paper verifies the optimization ability of the HGATPGA algorithm in the TPG task. The Maximum Fitness Curves (MFC) of different algorithms are shown in Figure 8.

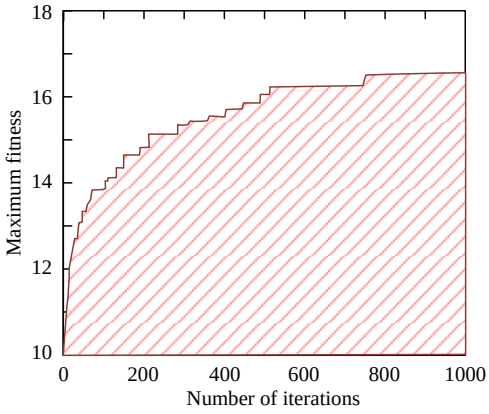
Figure 8 shows the population MFC maps of HGATPGA, AGA, and NGA. In Figure 8 (a), as the Number of Evolutionary Generations (EGN) increases, the Population Maximum Fitness Value (PMFV) of the research algorithm also increases continuously. When the EGN is about 400, the MFC of the research algorithm tends to stabilize, and the PMFV at this time is 18. When EGN is about 800, the PMFVs of AGA and NGA are 17 and 16, respectively. Compared with the traditional TPG, the research algorithm has a larger fitness value and can converge to a higher fitness level faster, indicating that the algorithm has better optimization ability and efficiency in solving related problems. To verify the model performance, this study conducts 20 comparisons with the two algorithms mentioned above, as exhibited in Figure 9.



(a) HGATPGA algorithm population maximum fitness curve

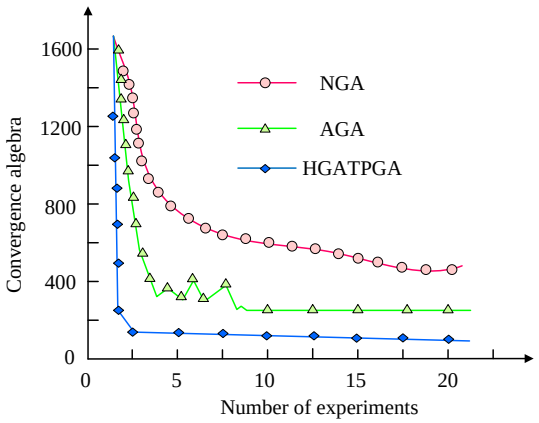


(b) AGA algorithm population maximum fitness curve

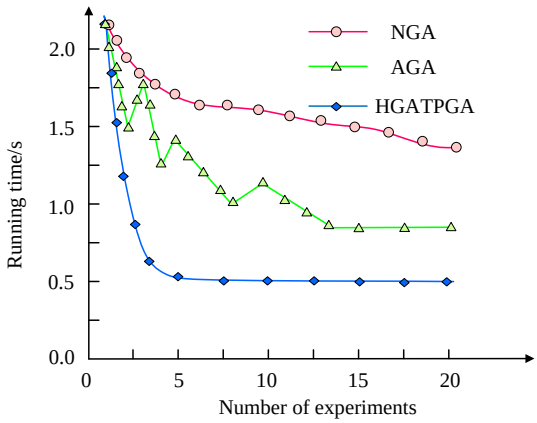


(c) NGA algorithm population maximum fitness curve

Figure 8: Population MFCs for different algorithms



(a) Convergence algebraic graphs of different algorithms



(b) Run-time graphs of different algorithms

Figure 9: Convergence algebra of different algorithms and graph of algorithm time

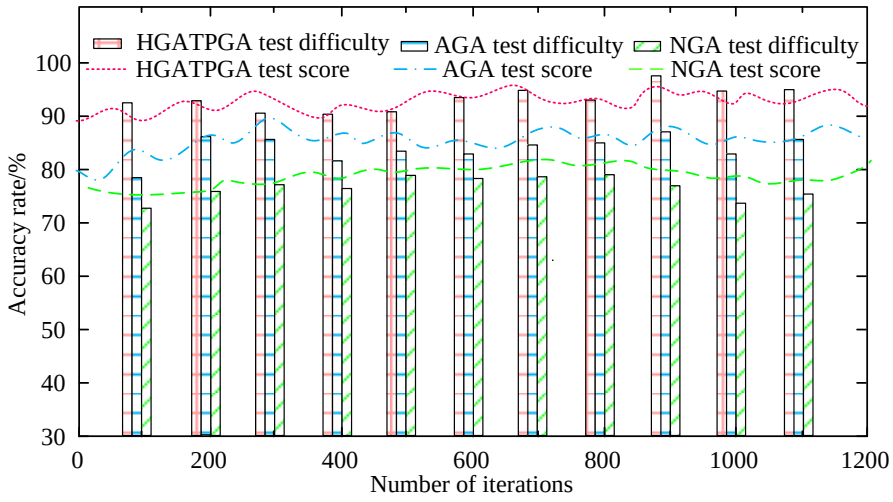
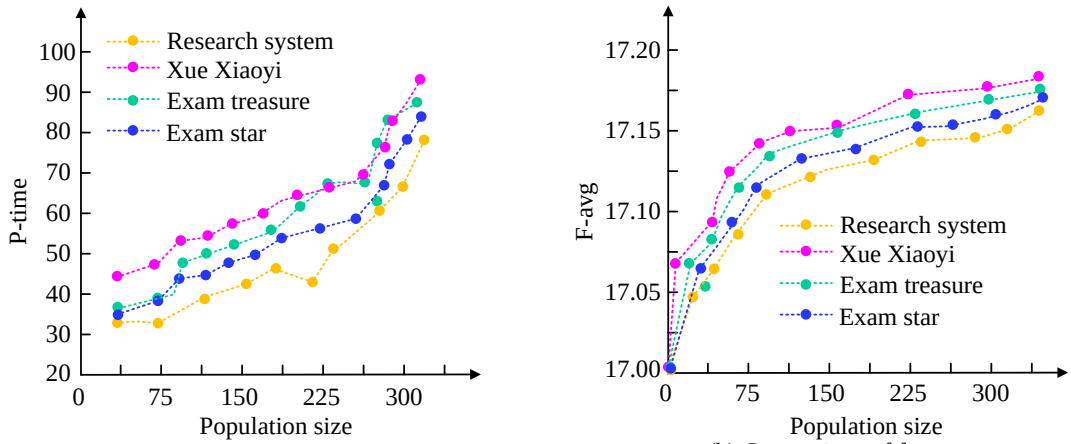


Figure 10: Comparison of the accuracy of different algorithms

Figures 9 (a) and (b) show the Convergence Algebra (CA) and running time of different algorithms. In the Experiments 1 to 20, the CA curve of HGATPGA tends to stabilize around 180 in the first experiment, and the fastest running time is 0.5s. The CA curves of AGA and NGA tend to stabilize around 320 and 480, with running times of 0.9s and 1.4s, respectively. Therefore, HGATPGA performs well in both convergence speed and runtime, making it the optimal choice among the three algorithms. It can achieve stable solutions in a short period and effectively improve computational efficiency.

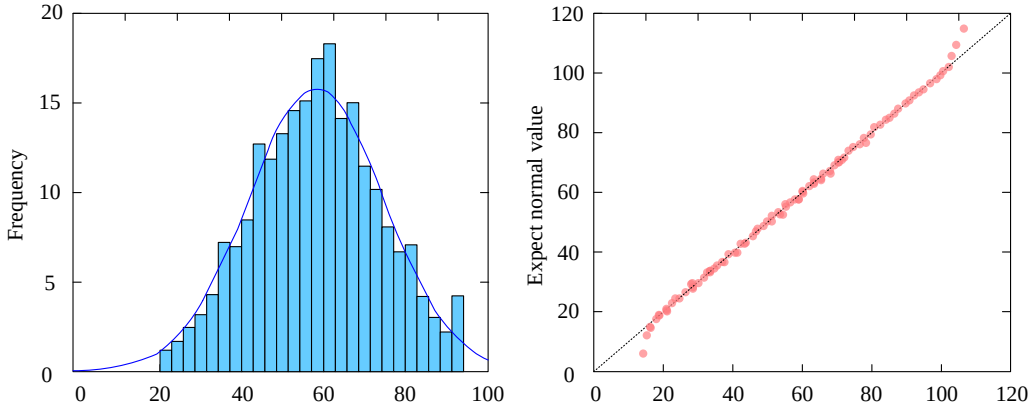
The accuracy rate of the difficulty test paper is indicative of the degree of correspondence between the generated test paper and the preset target in the difficulty distribution. The accuracy rate of the generated test paper is associated with the precision of predicting the score distribution of students. The research compares the accuracy of difficulty test paper grouping and the accuracy of score test paper grouping of HGATPGA, NGA, and AGA algorithms to comprehensively evaluate the performance of different algorithms in the task of TPG. The results are shown in Figure 10.



(a) Comparison of average test time of different online test systems

(b) Comparison of fitness mean values of different online test systems

Figure 11: Comparison of ATT and mean fitness of different OTSs



(a) Histogram of normal distribution of results of online test system

(b) Q-Q diagram of normal distribution of results of online test system

Figure 12: The normal distribution histogram and Q-Q diagram of the results of OTS

In Figure 10, the average accuracy of HGATPGA's Test Papers Difficulty (TPD) is as high as 92%, indicating its extremely high precision in ensuring a balanced distribution of TPD. In terms of TPS, the average accuracy of HGATPGA is 93%, further proving its strong ability in predicting student scores and optimizing the accuracy of TPS. The average accuracy of TPD for AGA is 81%, which is lower compared to HGATPGA, indicating that there may be some deviation or deficiency in difficulty control. In terms of TPS, the average accuracy of AGA is 85%, which is lower than HGATPGA, indicating that there may be some errors in predicting the distribution of student scores.

The average accuracy of NGATPD and TPS is relatively low, at 75% and 78%, indicating that the accuracy of NGA in TPD and score prediction needs to be improved.

In summary, there are significant differences in the accuracy of TPD and TPS among the three algorithms, with HGATPGA showing the highest test paper composition. The statistical verification results of the three algorithms are shown in Table 2. Table 2 shows that the three algorithms have extremely significant differences in CA and run time ($p<0.01$), with significant differences in difficulty and accuracy of generating grouped papers ($p<0.05$). HGATPGA is significantly superior to AGA and NGA in terms of convergence speed, run time, difficulty test accuracy, and group test accuracy. Significance test results further confirm the statistical significance of these differences, indicating that HGATPGA is the best choice among the three algorithms.

Table 2: Statistical verification of the three algorithms

Index	Pop Size	Constraint Level	HGATPGA	AGA	NGA	Significance
Convergence algebra	100	High	180±10	320±20	480±30	p<0.01
	200	Middle	170±15	310±25	470±35	p<0.01
	300	Low	160±10	300±20	460±30	p<0.01
Run time (s)	100	High	0.50±0.05	0.90±0.10	1.40±0.15	p<0.01
	200	Middle	0.48±0.04	0.88±0.08	1.38±0.12	p<0.01
	300	Low	0.46±0.03	0.85±0.07	1.35±0.10	p<0.01
Accuracy rate of difficulty test	100	High	92%±2%	81%±3%	5%±4%	p<0.05
	200	Middle	91.5%±1.5%	80%±2.5%	74%±3.5%	p<0.05
	300	Low	91%±1%	79%±2%	73%±3%	p<0.05
The grouped volume accuracy is obtained	100	High	93%±1.5%	85%±2.5%	78%±3.5%	p<0.05
	200	Middle	92.5%±1%	84%±2%	77%±3%	p<0.05
	300	Low	93%±0.5%	83%±1.5%	76%±2.5%	p<0.05

Table 3: Experimental environment setting parameters

Experimental tool	Specification parameter
Operating system	Windows 7 flagship 64-bit operating system
Install Memory	6.00 GB
processor	Intel(R)Core (TM)i5-2410M CPU@2.30GHz
Hard disk	463.00 GB
Development tool	Eclipse Juno SR2, jdk 1.8.0, tomcat 8.0
Archive	PostgreSQL 9.4
Language	Java
Graphics Card	NVIDIA GeForce GT 540M (1 GB VRAM)
Motherboard	ASUS K53E Series (Intel HM65 Express Chipset)
Display Resolution	1366 x 768 pixels
Network Interface	10/100/1000 Mbps Ethernet, Wi-Fi 802.11 b/g/n

3.2 Analysis of OTS results based on HGATPGA

Environmental configuration has a significant impact on the accuracy and reliability of experimental results. To ensure the consistency and reproducibility of the experiment, the relevant environment is set up in this study, as listed in Table 3.

This study selects the three OTSs with the highest download rates, namely Xue xiaoyi, Exam Treasure, and Exam Star, and compares them with the proposed OTS. Figure 11 compares the Average Test Time (ATT) and fitness mean of different systems.

In Figure 11 (a), among 300 population sizes, the ATT of the research system is the fastest, at 78.38s. The ATT of Exam Treasure (86.26s) is higher than that of the research system, while the ATT of Exam Star (81.42s) is higher than that of Exam Treasure. Xue Xiaoyi's ATT is the slowest, at 90.35 seconds. In Figure 11 (b), when the individual population size is 300, the mean fitness of the research system is 17.173, Exam Treasure is 17.162, Exam Star is 17.158, and Xue Xiaoyi has the lowest mean fitness at 17.153. To verify the accuracy of OTS scores, this study conducts a normal distribution test on OTS scores using SPSS tools. The distribution histogram and Q-Q plot are displayed in Figure 12. In the histogram of Figure 12 (a), the horizontal axis represents the distribution of scores, and the vertical axis represents the frequency of each score segment. Among them, the histogram distribution trend of OTS scores closely follows the normal distribution curve, indicating that the system's score distribution has good normality. In the Q-Q plot of Figure 12 (b), the data points obtained by OTS fluctuate closely around a straight line, indicating that the scores obtained by OTS are highly consistent with the theoretical normal distribution data. This confirms that the system's scores generally follow a normal distribution and have statistical significance.

4 Discussion and conclusion

4.1 Discussion

In the context of educational informatization, the automatic TPG system, as an important tool for improving teaching efficiency and quality, has attracted much attention for its performance and accuracy. This study proposes HGATPGA by integrating DL technology into the design of automatic TPG algorithm to improve the performance of OTS.

From the comparative analysis of algorithm performance, in terms of running time, the average run time of HGATPGA is only 0.5 seconds, which is much lower than that of AGA (0.9 seconds) and NGA (1.4s), indicating that HGATPGA has significant advantages in reducing computing costs and improving computing efficiency. This is the same as the results obtained by Rodriguez-Alvarez et al. [20]. The accuracy rate of test paper grouping is one of the important indexes to measure the performance of automatic test paper grouping system. The average accuracy rate of difficult test paper grouping by HGATPGA is 92%, and the average accuracy rate of test paper grouping is 93%, both of which reach a very high level. However, AGA and NGA have obvious shortcomings in the accuracy of test paper composition, especially in difficulty control and score prediction, indicating that HGATPGA has stronger ability in ensuring the quality of test paper, balancing the difficulty distribution and accurately predicting students' scores. This is consistent with the conclusion obtained by Li's team in the study of high-flux synthesis of binary Al-Mn alloy based on directed energy deposition [21]. The high performance of HGATPGA is mainly due to the integration of its DL module, and Transformer architecture can efficiently extract complex features in the test data to provide more accurate input information for the optimization algorithm. Through the combination of DL and GA, HGATPGA achieves the balance of global search ability and local search ability, so it is superior to AGA and NGA in terms of convergence speed and accuracy. Despite the enhancements made to AGA and NGA based on traditional GAs, they exhibit deficiencies in deep mining complex features. This limitation leads to sub-optimal performance in terms of optimization efficiency and accuracy.

In the comparison of ATT and mean fitness of OTS, under the condition of 300 population sizes, the proposed OTS performed the best on ATT, only 78.38s, significantly lower than other systems. Meanwhile, as the population size increased, the mean fitness of the system also continued to rise. When the population size was 300, the mean fitness reached 17.173, which was higher than other comparison systems. Therefore, the online test system proposed in the study has excellent performance in both test efficiency and fitness, and the results are consistent with those obtained by Yu et al. [22]. This study conducted a normal distribution test using SPSS tools, and the test results showed that the distribution trend of system scores closely followed the normal distribution curve, with each data point fluctuating closely around a straight line. This indicates that the system scores have good normality and statistical significance, proving the accuracy and reliability of OTS scores and providing strong support for educational evaluation.

In summary, the HGATPGA and its OTS designed in combination with DL technology in this article have shown excellent performance in convergence speed, run time, test paper accuracy, and score accuracy. Compared to traditional NGA and AGA, HGATPGA has higher optimization capabilities and efficiency, which can better meet the demand for high-quality test papers in education and teaching.

4.2 Conclusion

This study proposed a HGATPGA for OTS to address the issue of automatic TPG in computers. Performance testing was conducted on HGATPGA, with PMFVs of 18, 17, and 16 for research algorithm, AGA, and NGA. The average accuracy of TPD and TPS of HGATPGA was as high as 92% and 93%, respectively. Performance testing was conducted on OTS, and among 300 population sizes, the ATT (78.38s) of the research system

was the fastest, Exam Treasure (86.26s) was higher than the research system, and Exam Star (81.42s) was higher than Exam Treasure. Xue Xiaoyi's ATT was the slowest, at 90.35s. To sum up, the proposed online test can more accurately capture learners' knowledge state and individual needs by introducing DL technology, thereby generating test papers that are more in line with teaching objectives. The research focuses on the optimization of test paper composition in general scenarios, and the support for interdisciplinary integration and personalized learning needs is still insufficient. In the future, learners' portraits, knowledge point association graphs, and DL-driven adaptive recommendation technology can be combined to achieve a more accurate and personalized test composition scheme.

Funding

The research is supported by Guangxi University's Basic Skills Enhancement Program for Young and Middle-aged Faculty—Research on Intelligent Automated Grading Systems for Universities in the Context of Big Data (No. 2023KY1595).

References

[1] J. Li, "Improved Genetic Algorithm Enhanced with Generative Adversarial Networks for Logistics Distribution Path Optimization," Informatica, vol. 49, no. 11, pp. 67-82, 2025. https://doi.org/10.31449/inf.v49i11.6961

[2] M. Sang, Y. Ding, and M. Ding, "Metrics and quantification of power-line and pipeline resiliency in integrated gas and power systems," IET Generation, Transmission & Distribution, vol. 15, no. 21, pp. 3001-3016, 2021. https://doi.org/10.1049/gtd2.12236

[3] B. I. Bitachon, M. Eppenberger, and B. Baeuerle, "Reducing training time of deep learning based digital backpropagation by stacking," IEEE Photonics Technology Letters, vol. 34, no. 7, pp. 387-390, 2022. https://doi.org/10.1109/LPT.2022.3162157

[4] K. Simon, M. Vicent, K. Addah, D. Bamutura, B. Atwiine, D. Nanjebe, and A. O. Mukama, "Comparison of deep learning techniques in detection of sickle cell disease," Artificial Intelligence and Applications, vol. 1, no. 4, pp. 252-259, 2023. https://doi.org/10.47852/bonviewAIA3202853

[5] M. A. Lundine, L. L. Brothers, and A. C. Trembanis, "Deep learning for pockmark detection: Implications for quantitative seafloor characterization," Geomorphology, vol. 421, no. 15, pp. 1-20, 2023. https://doi.org/10.1016/j.geomorph.2022.108524

[6] B. I. Bitachon, M. Eppenberger, and B. Baeuerle, "Reducing training time of deep learning based digital backpropagation by stacking," IEEE Photonics Technology Letters, vol. 34, no. 7, pp. 387-390, 2022. https://doi.org/10.1109/LPT.2022.3162157

[7] Y. Wang, "Deep Learning Models in Computer Data Mining for Intrusion Detection," Informatica, vol. 47, no. 4, pp. 555-568, 2023. https://doi.org/10.31449/inf.v47i4.4942

[8] I. Latreche, S. Slatnia, and O. Kazar, "A Review on Deep Learning Techniques for EEG-Based Driver Drowsiness detection systems," Informatica, vol. 48, no. 3, pp. 359-378, 2024. https://doi.org/10.31449/inf.v48i3.5056

[9] Z. Yuan, "Interactive intelligent teaching and automatic composition scoring system based on linear regression machine learning algorithm," Journal of Intelligent and Fuzzy Systems, vol. 40, no. 2, pp. 2069-2081, 2021. https://doi.org/10.3233/JIFS-189208.

[10] L. Zhao, X. Zeng, and H. Ma, "Research and realization of automatic test system for motor controller unit," EDP Sciences, vol. 4, no. 13, pp. 1-9, 2021. https://doi.org/10.1051/e3sconf/202126003001

[11] R. Li, "Computer embedded automatic test system based on VxWorks," International Journal of Embedded Systems, vol. 15, no. 3, pp. 183-192, 2022. https://doi.org/10.1504/IJES.2022.124839

[12] J. Zhao, and J. Chen, "Automatic scoring system for CET-4 compositions based on Seq2Seq+Bi-LSTM Model," IEEE, vol. 3, no. 7, pp. 333-336, 2020. https://doi.org/10.1109/AUTEEE50969.2020.9315678

[13] M. Jelsch, Y. Roggo, A. Mohamad, "Automatic system dynamics characterization of a pharmaceutical continuous production line," European journal of pharmaceutics and biopharmaceutics: Official journal of Arbeitsgemeinschaft fur Pharmazeutische Verfahrenstechnik e.V, vol. 180, pp. 137-148, 2022. https://doi.org/10.1016/j.ejpb.2022.09.010

[14] B. Sitkowska, M. Kolenda, and D. Piwczyński, "Comparison of the fit of automatic milking system and test-day records with the use of lactation curves," Animal Bioscience, vol. 33, no. 3, pp. 408-415, 2020. https://doi.org/10.5713/ajas.19.0190

[15] L. Yan, C. Long, and W. Xie, "The hierarchy of a generalized automatic test system for electrical signals," 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), vol. 7, no. 3, pp. 1-5, 2020. https://doi.org/10.1109/ITNEC48623.2020.9084848

[16] I. H. Choi, and E. M. Kim, "Automatic construction of road lane markings using mobile mapping system data," Sensors and Materials: An International Journal on Sensor Technology, vol. 34, no. 7 Pt.2, pp. 2625-2635, 2022. https://doi.org/10.18494/sam3872

[17] J. Y. Wang, Y. H. Yin, and J. Y. Zheng, "Least absolute shrinkage and selection operator-based prediction of collision cross section values for ion mobility mass spectrometric analysis of lipids," Analyst, vol. 147, no. 6, pp. 1236-1244, 2022. https://doi.org/10.1039/d1an02161c

[18] Z. Wang, L. Wang, and Q. Jiang, "Multiple search

operators' selection by adaptive probability allocation for fast convergent multitask optimization," The Journal of Supercomputing, vol. 80, no. 11, pp. 16046-16092, 2024. https://doi.org/10.1007/s11227-024-06016-w

[19] M. Kojima, "Variable selection using inverse probability of censoring weighting," Statistical Methods in Medical Research, vol. 32, no. 11, pp. 2184-2206, 2023. https://doi.org/10.1177/09622802231199335

[20] J. S. Rodriguez-Alvarez, P. Khooblall, H. Brar, D. Fedrigon, J. Gutierrez-Aceves, and M. Monga, "Endoscopic stone composition identification: Is accuracy improved by stone appearance during laser lithotripsy?" Urology, vol. 182, no. 6, pp. 67-72, 2023. https://doi.org/10.1016/j.urology.2023.09.025

[21] D. Li, Y. Wu, M. Song, C. Chen, and K. Zhou, "High-throughput synthesis of binary Al–Mn alloys by directed energy deposition with improved accuracy of composition adjustment," Journal of Materials Science, vol. 59, no. 22, pp. 10022-10034, 2024. https://doi.org/10.1007/s10853-024-09462-2

[22] Z. Yu, Z. Wang, C. Xu, X. Zhou, Z. Chen, and C. Ren, "Comprehensive physical commutation characteristic analysis and test of hybrid line commutated converter based on physics compact model of IGCT," IEEE Transactions on Power Electronics, vol. 38, no. 2, pp. 1924-1934, 2023. https://doi.org/10.1109/TPEL.2022.321349