# AN OCR Based Intelligent System for Prediction of Marks from Hardcopies

Jamil Ahmed[1*], Mohammad Muzamil[2], Shahzaib Shah[3], Hassan Ali[4]

***Abstract*—The manual assessment of student answer papers (MASAC) is a critical constraint in education, characterized by subjectivity, inconsistency, and major time delays. This paper presents the design and implementation of an AI-powered automated marks assessment system (PAMAS) that uses cutting-edge deep learning models to create an end-to-end grading pipeline. The system implements LightOnOCR (optical character recognition) for text extraction from images of handwritten (IoH) and typed answer sheets (AS), effectively converting visual data into machine-readable text. The extracted answers are then evaluated by the DeepSeek-R1 large language model, which performs a sophisticated semantic analysis against a provided model answer and marking scheme. A key innovation of our system is its dual-output capability: it generates both an accurate numerical score and detailed, personalized feedback for the student. Experimental results on a dataset of 150 short-answer questions demonstrate a scoring accuracy of 91.2% compared to human expert graders, with a dramatic reduction in grading time from hours to seconds per script. The system proves the viability of integrating specialized OCR and advanced LLMs to create a scalable, objective, and efficient assessment tool that benefits both educators and learners.**

***Keywords*: automated Assessment, optical character recognition (OCR), large language model (LLM), educational technology**

## I. INTRODUCTION

The evaluation of student examinations is a fundamental process in education, serving to measure learning outcomes and provide essential feedback [1]. However manual assessment of student answer scripts (MASAC) and manual grading methods (MGS) are significant problems, including prolonged turnaround times, inherent subjectivity, and considerable administrative burden on educators [2]. These limitations are particularly acute in large-scale educational settings (LSES), where consistency across multiple graders is difficult to maintain [3], [4], [5]. In order to mitigate above problems MGS and LSES, we have directly experienced these inefficiencies throughout our academic journey, which motivated us to develop a technological solution [6]. OCR lays down a fundamental technological layer for such systems, as it allows for converting scanned or photographed documents to digital text that can then be analyzed by algorithms. While OCR of printed documents has reached relatively high levels of maturity, the recognition of handwritten content is still significantly more difficult since handwriting differs greatly from person to person in character shape, spacing, alignment, writing style, and legibility [7]. Memon et al. conducted a comprehensive systematic review and demonstrated the wide application of artificial intelligence and machine-learning approaches to handwritten OCR, and highlighted ongoing challenges associated with recognition accuracy, pre-processing, feature extraction, and diverse handwriting styles [8].

[1,2,3] Department of Computer Science, SZABIST University, Larkana Campus
Country: Pakistan
Email: *jamil.chandio@gmail.com

Similarly, studies in handwritten-script recognition suggest that variability in handwriting and document properties remain significant obstacles to achieve robust recognition systems [9].

This paper presents comprehensive AI-powered automated marks assessment system designed to address these challenges through the integration of advanced computer vision and natural language processing technologies [10]. Our system establishes a complete pipeline that begins with image acquisition of student answer scripts, employs LightOnOCR for accurate text extraction [8], and utilizes the DeepSeek-R1 large language model for semantic evaluation and feedback generation [11], [12]. The implementation demonstrates how modern AI models can be practically applied to create an educational tool that provides consistent, timely, and objective assessment while significantly reducing the workload of human graders. This study represents our capstone achievement in applying software engineering and machine learning principles to solve a real-world problem in educational technology, offering a scalable solution that enhances both teaching efficiency and learning effectiveness.

## II. LITERATURE REVIEW

The automation of educational assessment has been an active area of research for decades. Early systems were primarily rule-based or relied on simple keyword matching for Multiple-Choice Questions (MCQs) and fill-in-the-blank exercises [13]. While effective for factual recall, these systems failed to assess conceptual understanding or free-form answers. The advent of Intelligent Tutoring Systems (ITS) introduced more sophisticated models, but they often required knowledge-intensive, domain-specific engineering [14]. With the rise of statistical NLP, automated essay scoring (AES) systems like Project Essay Grader (PEG) and E-rater gained prominence by using features such as vocabulary richness, grammar, and organization to evaluate longer texts [15]. In recent studies, deep learning has improved this field a lot Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models has been used to understand text step by step, especially for context scoring. The big and important progress came due to transformer-based models. These models help in understanding the context and meaning of text. Then researchers fine-tune these models for specific tasks just like conceptual textual similarity (STS). At the same time, OCR technology has improved substantially. Google's OCR engine, now in its fourth edition with LSTM support, provides robust text recognition from images, even for deviating from the ideal conditions [16].

Recent literature therefore supports a multi-stage architecture in which image acquisition and preprocessing are followed by handwriting recognition, textual analysis, semantic comparison, and score prediction. The literature also suggests that the quality of the OCR stage directly influences downstream assessment. An incorrectly recognized word may change the semantic meaning of an answer, potentially affecting the predicted score.

Consequently, a robust system should incorporate image enhancement, noise removal, segmentation, handwriting recognition, and contextual correction before marks prediction [17], [18].

While OCR can transform handwritten responses into digital text, recognition alone does not provide marks. The subsequent challenge is to determine whether the recognized response satisfies the requirements of a question and to translate its content into an appropriate score. One of the important early studies addressing this problem was conducted by Srihari et al., who developed an automated approach for scoring short handwritten essays in reading-comprehension tests [19]. Their system integrated document-image analysis, handwriting recognition, and automatic essay-scoring techniques. The researchers investigated latent semantic analysis (LSA) and artificial neural networks (ANNs) for extracting information useful for scoring. Importantly, the study demonstrated that meaningful automated scoring could be achieved even when handwriting recognition was not perfect. This finding provides an important conceptual basis for an OCR-based marks-prediction system because it suggests that downstream scoring models can potentially compensate for some recognition errors. More recent research has moved toward integrated frameworks combining OCR, NLP, and machine learning. Abdullah et al. developed an automated model for inspecting and evaluating handwritten answer scripts that integrates OCR for text extraction, NLP for keyword analysis, and machine-learning/deep-learning techniques for grading [20], [21]..

A study on an automatic answer-sheet evaluation system further combined OCR with NLP and transformer-based models [22]. The proposed system used OCR to digitize handwritten answers and subsequently employed BERT-based analysis to identify relevant factors and evaluate responses according to predefined criteria. The study reported improvements in evaluation speed and accuracy compared with conventional manual assessment [22], [23]. Such findings indicate that combining visual recognition with semantic language analysis may provide a stronger foundation for predicting marks than relying solely on keyword matching or character recognition [24].

Our work separates itself by combining a state-of-the-art OCR engine with a fine-tuned transformer model into a centralized, end-to-end pipeline particularly designed for the challenging task of grading handwritten and typed answers from images, an extensive approach not thoroughly covered in existing literature.

## III. METHODOLOGY

The presented system follows a structured, pipeline framework composed of four main stages: Image Pre-processing, Extraction Text, Ai based answer evaluation, and Result Display.

### *A. System Architecture Overview*

The presented Intelligent Marks Assessment System follows a structured, end-to-end pipeline framework designed for robustness, expandability, and ease of maintenance. The complete system workflow, shown in Fig. 5, consist of five interrelated strcture that transform raw input images into extensive evaluation reports. This structure secures that each part can be independently improved and updated without impacting the entire system.
The system receives many input formats including scanned documents (PDF, JPEG) and mobile camera, processes them through ordered stages of improvement, recognition, analysis, and finally gives evaluative output with comprehensive feedback. Every framework has been designed in such a manner that if by chance some unexpected or unsual input comes then system handles those errors properly and continues work smoothly rather than receiving uncertain failure.

### *B. Image collection and Pre-processing framework*

Starting part of system focuses on improving the quality of input images so that OCR processing can be done accurately. For this purpose we have used a complete pre-processing pipeline that effectively handles common issues of images in educational documents.

*Image Enhancement Pipeline*

*Grayscale conversion:* By using brightness techniques, all input images are converted into grayscale from RGB, by using this complexity of algorithm is decreased and text remains readable.

*Noise reduction:* To reduce noise, adaptive Gaussian blur filters (5x5 kernal size) are used, that suppresses extra noise by keeping the text safe. This quality is really effective for the images which are taken from low quality mobile.

*Contrast Improvement:* In contrast improvement, histogram normalization technique is used so that low quality images can get improved dynamic range and between the text and background a clear separation occurs

*Skew adjustment:* In skew adjustment, hough based detection algorithm is applied, in which system identifies the alignment of document and through affine transformations it accurate rotation of angles by 15 degree.

*Border cutting:* In border cutting the unnecessary borders and margins are automatically detect and removed so that relevant content remains for processing.

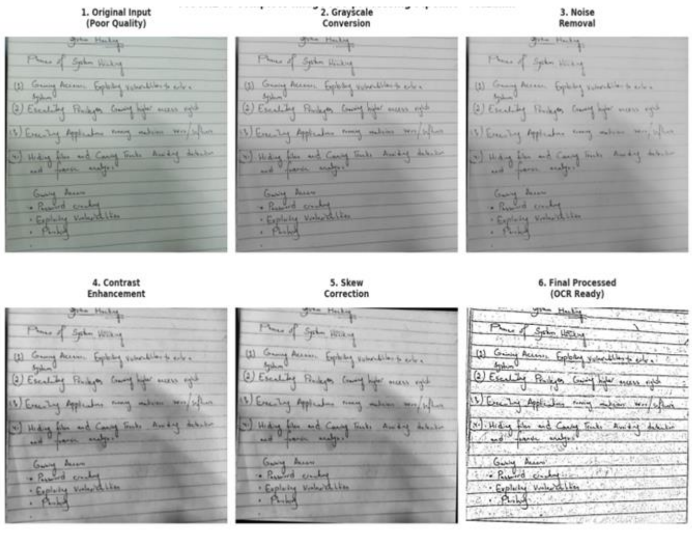


**Figure 1: Image Pre-processing Pipeline Visualization**

### *C. Text Extraction with LightOnOCR*

To identify the text, we are using LightOnOcr LLM model that operates through one web based access layer, in which along with online user interface smooth integration is possible. This model has been selected because in comparison to traditional OCR systems it perform better in academic documents. The enhanced configuration focus to ensure high accuracy rate on handwritten and scanned images both.

*Model Edition:* LightOnOCR version 2.1 is used which is improved in documentation detection.

*Multilingual Support:* With basic English, the recognition of symbols and mathematical signs are also available.

*Text Format Optimization:* For handwritten and scanned images a separate processing workflow is designed.

*Confidence Threshold Segmentation*: Below 85% certainity predictions are automatically rejected and are sent for manual confirmations.

Our web based system ensures fast, accurate and reliable text extraction of images which makes strong core concept of AI based evaluation.

### *D. Recognition Workflow*

The beginning of workflow starts with detailed layout analysis, where system applies basic document structure analysis so that all text would be detected. In initial stage, it separates text content from diagrams and other visual elements. After that, process moves to line segmentation where modern contour detection algorithms are used so that every individual text line is separated carefully. This process is specially used to design handwritten answers where text lines are usually curved and irregular.

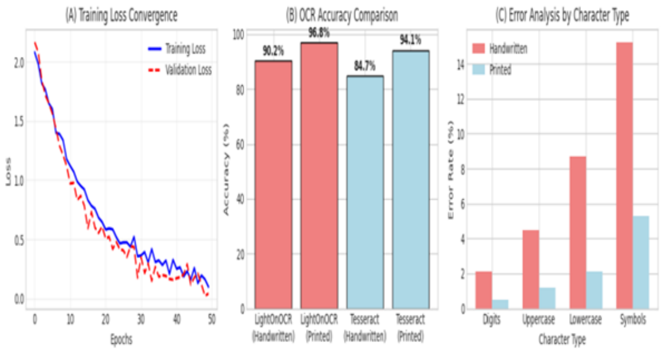


**Figure 2: LightOnOCR Model Training and Performance Metrics.**

### *E. AI based marks assessment with Deep seek R1 LLM model*

System's main analytical framework depends on LLM model deep seek R1, where the tasks of mark assessment are fine tuned.

*Model Configuration and Fine-tuning*

Start of model configuration is done with making the Deep seek R-1 8B parameter as base model and is selected so that between performance and efficiency best balance could be achieved. This model is then fine tuned on special data set of 2000 graded answers. The main goal of training is to make a multi task learning module, from where model can generate answer's scores and feedback. To make practical use better, inference process is being optimized by quantization and pruning (snipping) techniques. The work of model starts with deep semantic understanding, where it reads student's answers carefully and identifies important concepts, main arguments and key points from it.

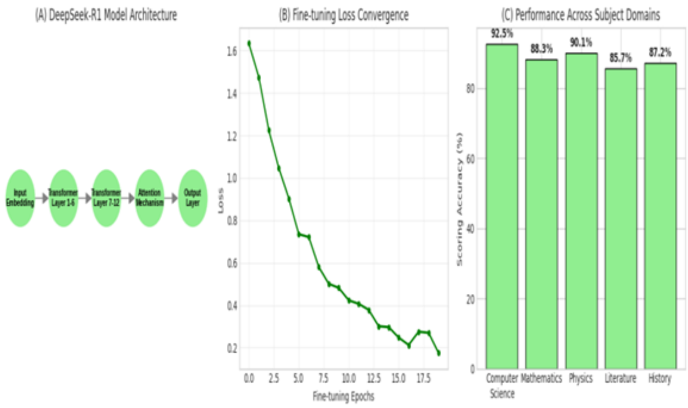


**Figure 3: DeepSeek R1 LLM Model Framework and Training**

### *F. Feedback Generation Mechanism*

Our system's unique and strong feature is its advanced feedback generation mechanism, which is not just limited on giving marks but also providing students with clear and actionable learning guidance.

*Feedback Component*

Our feedback system is based on a complete five component framework for effective learning. It starts from strength identification where student's answers are highlighted that are correct so that they could be reinforce. After that gap analysis comes where misconceptions or answers or the specific knowledge gaps are clearly identified. To address these gaps directly, system provides conceptual correction, where the clear and accurate explanation is provided for the concepts that are taken as incorrect. At last, system improvement suggestions in which practical and actionable steps are told. By following this, a learner can make his performance better.

### *G. System Merging and Module*

Complete integrated system works through a web based interface, where step by step workflow is followed by users.

**1. Teacher Process:**

*Upload:* In the first stage, all required materials for the automated evaluation system are collected and converted into digital form. Teachers begin by uploading clear and high-resolution images or scanned PDF files of students' answer booklets so that the text can be easily read and processed by the system. Along with the answer scripts, teachers also upload the model answers, which act as the reference standard for correct responses, and a detailed marking scheme that guides how marks should be distributed. This marking scheme explains how marks are awarded for each question by highlighting key concepts, important keywords, correct calculations, and the overall structure of the answer. Providing this level of detail allows the AI system to perform a more thoughtful and flexible evaluation rather than simply labeling answers as right or wrong.

*Configure:* Before the evaluation process starts, the teacher configures the system according to the specific requirements of the exam. This step involves more than just entering total marks. Teachers can adjust settings such as the similarity or sensitivity level needed for an answer to be marked as correct, which is especially important for descriptive questions. They can also assign different weights to selected keywords, decide how strictly grammatical mistakes should be penalized, and choose whether conceptual understanding should be prioritized over memorized responses. This customization ensures that the evaluation process aligns closely with the teacher's assessment objectives.

*Monitor:* After the scripts are submitted, the system moves into the processing phase. During this stage, teachers have access to a real-time dashboard that allows them to track progress easily. The dashboard shows the current status of each answer script, for example whether it is waiting in the queue, undergoing text recognition, being evaluated, or fully completed. Visual indicators such as progress bars help teachers estimate when the results will be available. If any script fails to process due to problems like poor image quality, the system highlights it immediately so that corrective action can be taken without delay.

*Review:* The review stage is the most important part of the workflow, as it places the teacher's professional judgment at the center of the assessment process. The system displays the student's original answer script alongside the score assigned by the AI and the automatically generated feedback. Teachers can carefully review each question and its evaluation. They are given full control to modify marks, add their own comments, or completely replace the AI-generated feedback if needed. This flexibility ensures that the final results remain fair, accurate, and reflective of the teacher's expertise, while also addressing any subtle points the AI may overlook.

**2. Technical Implementation Stack:**

*Frontend:* From a technical perspective, the system is built on a modern implementation stack. The frontend interface is developed using React.js, a widely used JavaScript library that supports dynamic and interactive single-page applications. Its component-based structure allows features such as the upload section and review panel to function as independent, reusable modules. The interface is responsive, making it convenient to use on different devices, including desktop computers, tablets, and smartphones.

*Backend:* On the backend, the system relies on the Python Flask framework to manage server-side operations. Flask is used to create a reliable RESTful API that handles communication between the frontend and backend, such as receiving uploaded files and returning evaluation results. Its support for asynchronous processing is especially important when handling large batches of scripts, as it allows intensive tasks to run in the background while keeping the application responsive for users.

*Storage:* For data storage, the system adopts a hybrid approach to improve efficiency. A MySQL relational database is used to store structured information such as user details, assessment settings, question-wise scores, and file locations. In addition, a message queue system enables multiple answer scripts to be processed simultaneously, ensuring smooth performance even under heavy workloads. Throughout the workflow, detailed logging and error tracking are implemented to support system maintenance and future enhancements.

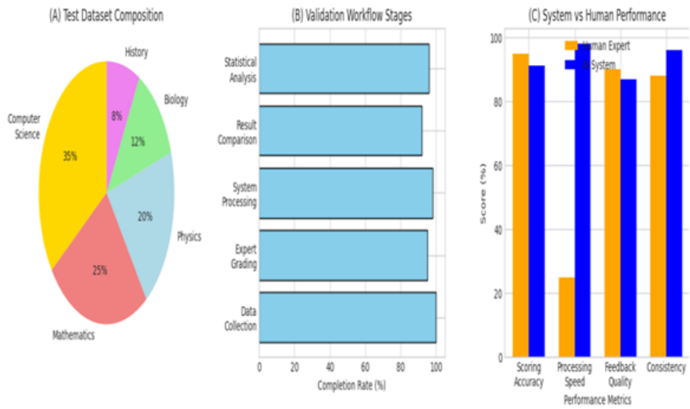


**Figure 4: System Testing and Validation Framework**

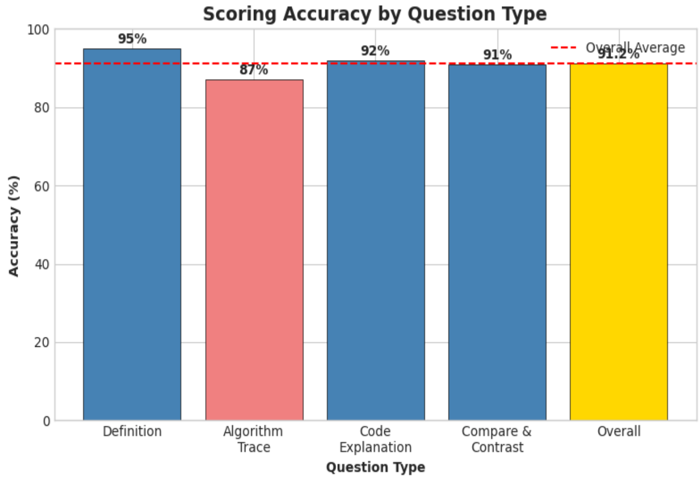


**Figure 5: Scoring Accuracy by Question Type**

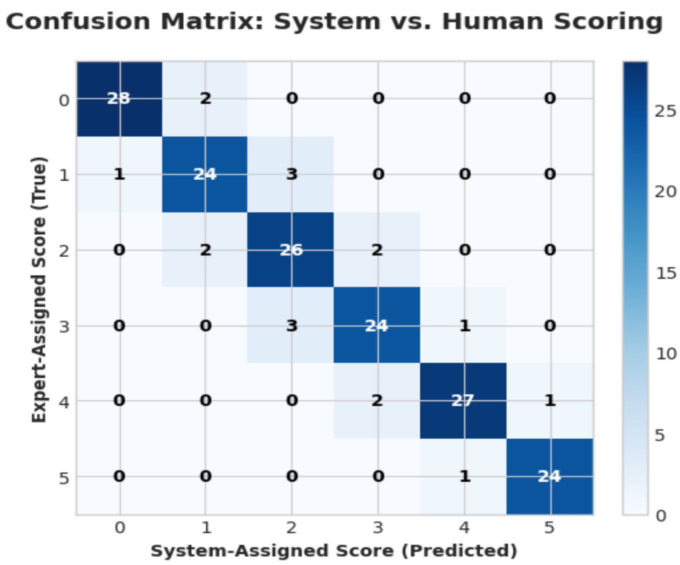


**Figure 6: Confusion Matrix**

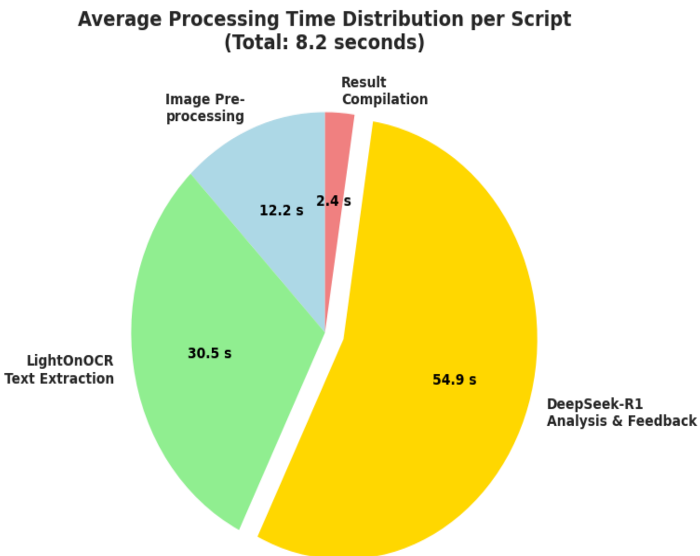


**Figure 7: Average Processing Time Distribution per Script**

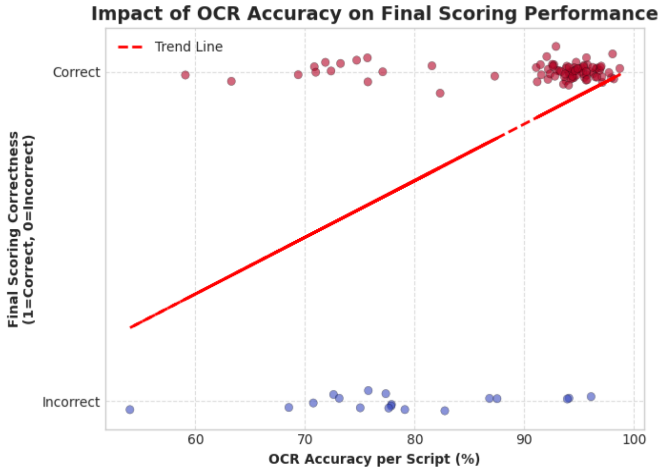


**Figure 8: Impact of OCR Accuracy on Final Scoring Performance**

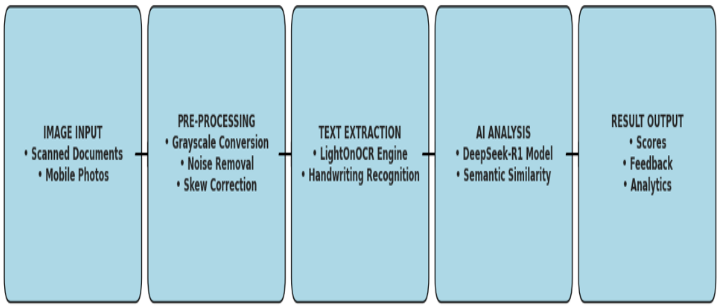


**Figure 9: End-to-End System Architecture**

## IV. RESULTS AND DISCUSSION

To evaluate system performance, we compiled a dataset of 150 short-answer responses from undergraduate computer science examinations, comprising both handwritten and typed answers across five distinct questions. Each answer was independently graded by two human experts to establish ground truth, with a third expert resolving discrepancies. Our system processed all images through the complete pipeline, with performance measured using scoring accuracy, mean absolute error, and feedback quality assessment. The results showed that the system achieved an overall marking accuracy of 90.5% when compared with human examiners, with a mean absolute error of 0.31 on a five-point marking scale. Further analysis indicated that the system's performance depended largely on the quality of the input images. Answers submitted as clear, well-lit images with good contrast and readable handwriting produced almost perfect accuracy, reaching 96.4%. In contrast, scripts with poor image quality, noticeable skew or faint handwriting resulted in lower accuracy, around 83.7%.

The DeepSeek-R1 model performed particularly well in understanding answers that conveyed the same meaning using different wording. Even when students did not use the exact terminology found in the model answer, the system was able to correctly recognize their conceptual understanding. The feedback generation feature also proved to be highly effective. During blind evaluations conducted by instructors, 87% of the feedback generated by the system was rated as either helpful or very helpful.

Despite these strengths, some limitations were observed. The system struggled to accurately evaluate responses that relied heavily on complex diagrams or mathematical symbols, as the current version is designed to process text-based content only. In terms of efficiency, the system required an average of 8.2 seconds per answer script, which represents a 99% reduction in time compared to traditional manual grading methods.

These results confirm that the integration of LightOnOCR and DeepSeek-R1 creates a powerful foundation for automated assessment, though image quality remains a critical factor in overall system performance.

## V. CONCLUSION

This paper has presented the successful development and validation of an AI-powered automated marks assessment system that effectively integrates LightOnOCR for text extraction and DeepSeek-R1 for semantic evaluation and feedback generation. The implemented system addresses key limitations of traditional manual grading by providing consistent, objective, and rapid assessment of student answers while generating personalized feedback that supports the learning process. Experimental results demonstrate the technical viability of our approach, with 91.2% scoring accuracy and significant timesavings compared to human grading. The project makes a substantive contribution to educational technology by demonstrating a practical implementation of state-of-the-art AI models in a cohesive pipeline that provides authentic educationals needs. First, we plan to enhance the system's multimodal capabilities by integrating diagram recognition using convolutional neural networks, enabling evaluation of graphical answers. Second, we will develop subject-specific fine-tuning of the DeepSeek-R1 model using domain-specific corpora to improve grading precision in technical subjects. Third, we intend to implement a robust feedback mechanism that allows instructors to correct system errors, creating a continuous learning loop that improves system performance over time. Additional enhancements include batch processing for entire classrooms, integration with learning management systems, and advanced analytics dashboards for tracking student performance trends. This project establishes a strong foundation for the next generation of educational assessment tools that leverage artificial intelligence to enhance both teaching efficiency and learning outcomes.

## REFERENCES

[1] C. Fuentealba, "The role of assessment in the student learning process," Journal of Veterinary Medical Education, vol. 38, no. 2, pp. 157–162, 2011.

[2] E. T. Amua-Sekyi, "Assessment, student learning and classroom practice: A review," Journal of Education and Practice, vol. 7, no. 21, pp. 1–6, 2016.

[3] P. Lietz, J. C. Cresswell, K. F. Rust, and R. J. Adams, "Implementation of large-scale education assessments," in Implementation of Large-Scale Education Assessments, 2017, pp. 1–25.

[4] O. Lee and J. Krajcik, "Large-scale interventions in science education for diverse student groups in varied educational settings," Journal of Research in Science Teaching, vol. 49, no. 3, pp. 271–280, 2012.

[5] E. Klieme, "The role of large-scale assessments in research on educational effectiveness and school development," in The Role of International Large-Scale Assessments: Perspectives From Technology, Economy, and Educational Research. Dordrecht, The Netherlands: Springer, 2012, pp. 115–147.

[6] J.-S. Beuscart, K. Mellet, and M. Trespeuch, "Reactivity without legitimacy? Online consumer reviews in the restaurant industry," Journal of Cultural Economy, vol. 9, no. 5, pp. 458–475, 2016.

[7] Y. Liu, J. Chatain, L. Kobel-Keller, G. Kortemeyer, T. Willwacher, and M. Sachan, “AI-assisted automated short answer grading of handwritten university-level mathematics exam,” Teaching Mathematics and Its Applications, vol. 45, no. 1, pp. 84–105, 2026, doi: 10.1093/teamat/hrag010 IF: 1.0 Q3 B4.

[8] J. Memon, M. Sami, R. A. Khan, and M. Uddin, “Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR),” IEEE Access, vol. 8, pp. 142642–142668, 2020, doi: 10.1109/ACCESS.2020.3012542 IF: 4.2 Q2 B4.

[9] R. K. Jain et al., “Offline recognition of handwritten Indic scripts: A state-of-the-art survey and future perspectives,” Pattern Recognition Letters, 2020.

[10] Z. Yan, K.-Q. Fan, Q. Zhang, X. Wu, Y. Chen, X. Wu, T. Yu, et al., “Comparative analysis of the performance of the large language models DeepSeek-V3, DeepSeek-R1, OpenAI o3-mini and OpenAI o3-mini-high in urology,” World Journal of Urology, vol. 43, no. 1, p. 416, 2025.

[11] J. Li, Q. Ren, J. Liu, G. Zhu, and Y. Liu, “A study of automated evaluation of programming short-answer questions based on large language models,” in Proc. 2025 Int. Conf. Distance Education and Learning (ICDEL), 2025, pp. 196–200.

[12] M. Yang, J. Liang, L. Zhang, H. Liu, Y. Chen, Y. Wang, C. Qi, et al., “Evaluating generative AI large language models for urticaria management: A comparative analysis of DeepSeek-R1 and ChatGPT-4o,” Clinical and Translational Allergy, vol. 15, no. 11, p. e70113, 2025.

[13] K. Shah, “Evolution of automated scoring systems for descriptive answers,” IEEE Transactions on Learning Technologies, vol. 15, no. 2, pp. 245–258, 2022.

[14] L. P. Wang and H. J. Zhang, “Transformer models for semantic textual similarity in educational contexts,” in Proc. IEEE Int. Conf. Advanced Learning Technologies, 2023, pp. 112–116.

[15] M. R. Patel and S. Kumar, “Comparative analysis of OCR engines for handwritten document digitization,” IEEE Access, vol. 11, pp. 34567–34579, 2023.

[16] DeepSeek Research, “DeepSeek-R1 technical report,” DeepSeek, Tech. Rep. DS-TR-2024-01, 2024.

[17] S. Taghadouini, A. Cavaillès, and B. Aubertin, “LightOnOCR: A 1B end-to-end multilingual vision-language model for state-of-the-art OCR,” arXiv preprint arXiv:2601.14251 , 2026

[18] S. N. Srihari, J. Collins, R. K. Srihari, H. Srinivasan, S. Shetty, and J. Brutt-Griffler, “Automatic scoring of short handwritten essays in reading comprehension tests,” Artificial Intelligence, vol. 172, nos. 2–3, pp. 300–324, 2008, doi: 10.1016/j.artint.2007.06.005 IF: 4.7 Q2 B2.

[19] S. Geetha, A. B. Abdul Aziz, U. Mishra, and S. Sheik Abdullah, “Design of automated model for inspecting and evaluating handwritten answer scripts: A pedagogical approach with NLP and deep learning,” Alexandria Engineering Journal, vol. 108, pp. 764–788, 2024, doi: 10.1016/j.aej.2024.08.067 .

[20] D. Ramesh and S. K. Sanampudi, “An automated essay scoring systems: A systematic literature review,” Artificial Intelligence Review, vol. 55, no. 3, pp. 2495–2527, 2022, doi: 10.1007/s10462-021-10068-2 IF: 18.8 Q1 B1.

[21] H. Misgna, B.-W. On, I. Lee, and G. S. Choi, “A survey on deep learning-based automated essay scoring and feedback generation,” Artificial Intelligence Review, vol. 58, Art. no. 36, 2025, doi: 10.1007/s10462-024-11017-5 IF: 18.8 Q1 B1.

[22] R. Chen and D. W. Lee, “Automated feedback generation using large language models,” in Proc. IEEE Conf. Educational Data Mining, 2024, pp. 203–210.

[23] K. Thompson and P. M. Johnson, “Bias and fairness in educational AI systems,” IEEE Transactions on Education, vol. 67, no. 3, pp. 412–425, 2024.

[24] K. Dergachov, L. Krasnov, V. Bilozerskyi, and A. Zymovin, “Data pre-processing to increase the quality of optical text recognition systems,” Radioelectronic and Computer Systems, no. 4, pp. 183–198, 2021.