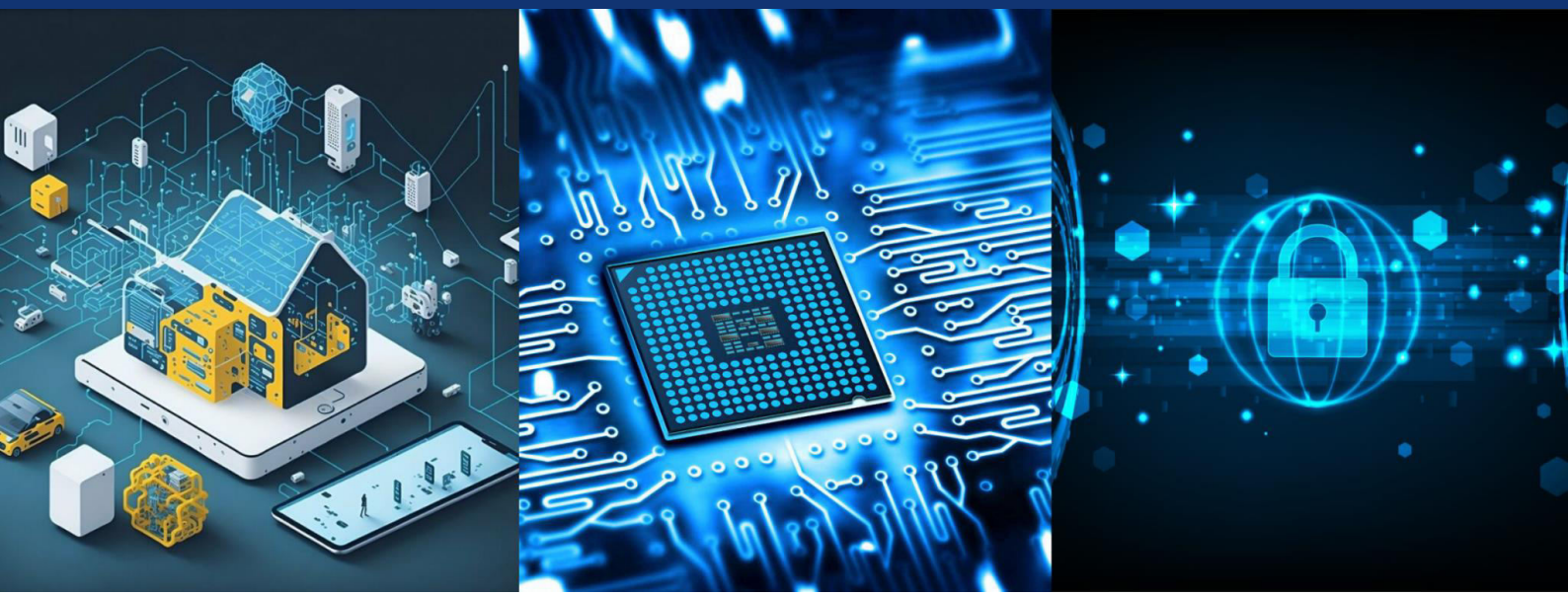
IJIRCCE

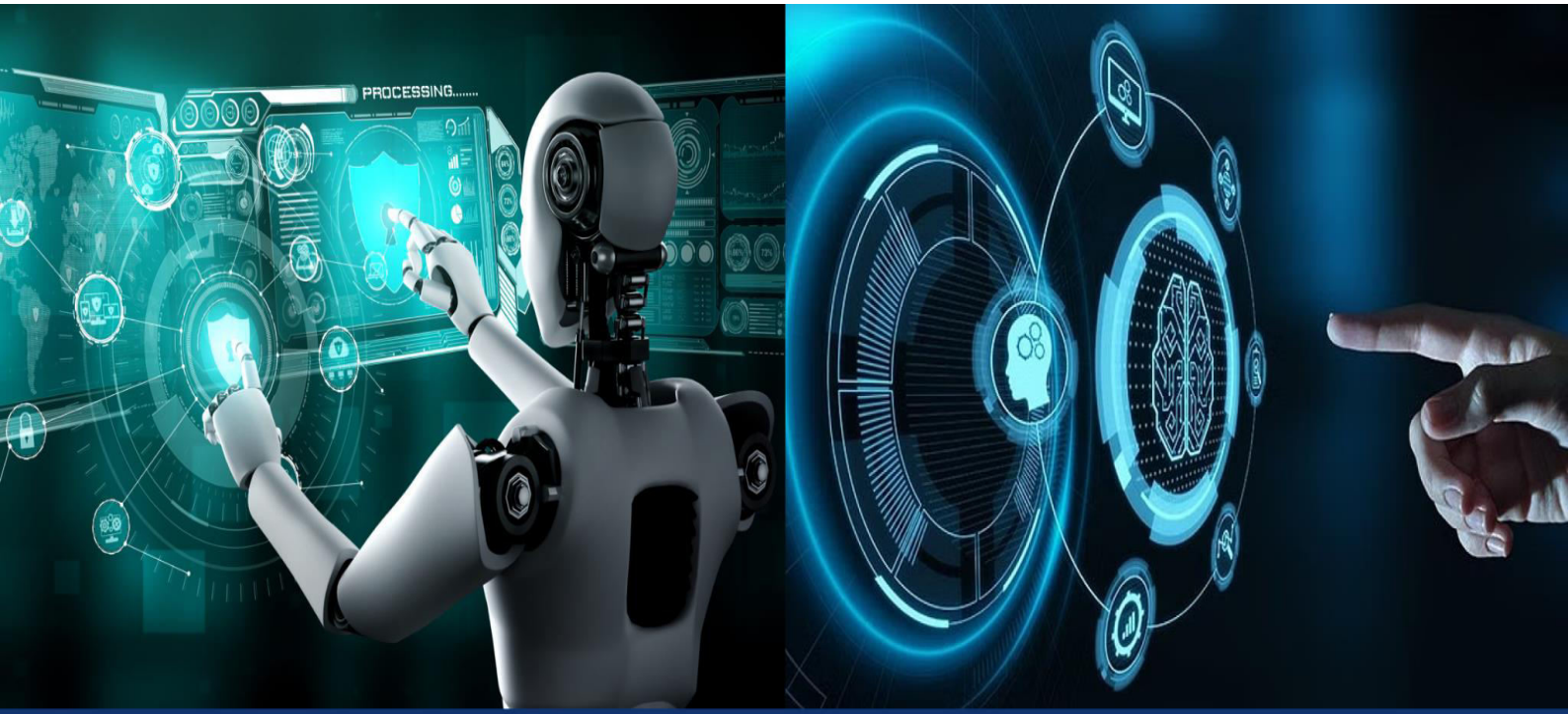
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 10, October 2026

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410051

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

AI-Based Handwritten Answer Sheet Evaluation Using OCR and NLP Semantic Analysis

Prof. Basavaraja Patil G V[1], Chaitra B S[2], Chandana S[2], Chandralekha M S[2], Harshita G K[2]

Assistant Professor, Dept. of CSE, Jain Institute of Technology, Davangere, Karnataka, India[1]

UG Students, Dept. of CSE, Jain Institute of Technology, Davangere, Karnataka, India[2]

ABSTRACT: Checking answer sheets that are written by hand is a slow and exhausting process, and because of evaluator fatigue and the need to make subjective judgments, it usually leads to inconsistent results—this is indeed a serious issue for organizations which have to evaluate a large number of students. This paper presents an AI-based system which automates the grading of descriptive answers by combining Optical Character Recognition (OCR) with a semantic analysis carried out using Natural Language Processing (NLP). The handwritten responses are first processed by a multimodal large language model, then improved using traditional NLP methods, and then compared to the correct answers using Sentence-BERT (SBERT) embeddings and cosine similarity. The similarity scores obtained in this way are converted into grades according to a grading rubric, and a full and detailed report is produced to ensure complete transparency. The system is based on the most recent advances in OCR, transformer-based similarity models, layout analysis, and LLM-assisted grading, and is designed to speed up the assessment process without missing what the student actually understood. The paper also examines open questions such as differences in handwriting, diagrams, mathematical notation, and validation against human graders, and identifies areas that need further research.

KEYWORDS: Optical Character Recognition (OCR), Natural Language Processing (NLP), Sentence-BERT (SBERT), Semantic Similarity, Automated Answer Evaluation, Large Language Models (LLMs), Handwritten Text Recognition, and Educational Assessment.

I. INTRODUCTION

Assessment is very important in the area of education since it indicates the extent to which students have understood the material, provides a record of what they have learned, and offers feedback that enables them to make improvements. However, despite the considerable advances that have been made in digital tools, the majority of answers written by hand are still assessed by humans. In large organisations, where it is necessary to examine hundreds or thousands of scripts in a short time, this approach places a heavy strain on the assessors: it occupies a great deal of their time, causes them to become tired, results in inconsistency, and leads to differences in the standards applied by different graders—all of which can cause two similar answers to be given different scores.

Advances in the fields of AI, computer vision, and natural language processing are altering the current situation. Nowadays, multimodal OCR models are able to recognize handwriting from a wide variety of styles much more reliably than earlier systems, and transformer-based language models such as BERT and its descendant SBERT can determine whether an answer communicates the correct idea even though its phrasing is entirely different from that of a model answer – a capability which keyword matching never had.

The study describes a system which combines these two features. First, the answer sheets are scanned and processed prior to the extraction of the handwritten text; the text thus obtained is normalised (by removing the noise, lemmatising it and splitting it into sentences) and is then compared with the model responses using SBERT embeddings and cosine similarity. The similarity score is included in a marking rubric that takes into account conceptual correctness, completeness and the inclusion of relevant keywords, and the system ultimately generates a report displaying the marks given for each question, the total score and the extracted text for review. The aim throughout is to reduce the time needed for grading, to improve the consistency of the marking, to minimise human bias and to enable scaling to large institutional volumes — just as in the case of evaluating answers on the basis of their meaning rather than on the exact wording, unlike the earlier keyword-matching tools.

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410051

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The paper is organised as follows: Section II looks at previous work on automated evaluation, OCR, semantic similarity, and the use of large language models for grading; Section III explains the architecture that has been proposed; Section IV sets out the method that will be used to assess the system; Section V examines its limitations; Section VI considers some possible future extensions; and Section VII presents the conclusion.

II. LITERATURE REVIEW

The method used to evaluate answers written by hand has changed greatly since initially involving just the scoring of keywords and is now based on the use of AI systems capable of understanding meaning and context. These advances have been made possible by developments in OCR, NLP, transformer models, and LLMs. The following section looks at this progress and points out the gaps that the system described in this paper aims to address.

A. Methods using traditional approaches and those using embeddings.
Initially, lexical techniques—such as keyword matching, Levenshtein distance, TF-IDF, and cosine similarity based on word counts—were employed in order to compare answers by examining the number of words they had in common. Although these methods worked well with short, objective answers, they performed poorly when the responses had been rephrased or included a large number of synonyms. To some extent, this problem was overcome by Word2Vec and GloVe, which represented words as vectors capable of capturing certain semantic relationships, while BERT went a step further by using context-aware embeddings. SBERT then enabled sentence-level embeddings to be produced for fast cosine-similarity comparisons, and various studies have shown that scoring using SBERT correlates more closely with the evaluations of human graders than lexical matching does—this is the reason why it has been selected as the semantic engine in this instance.

B. The identification of optical characters and the process of extracting handwritten text
The key issue is making sure that the text is captured accurately. Although traditional systems such as Tesseract perform well with printed text, they have a lot of trouble dealing with handwriting since variations in the shape of the letters, the spacing, and the quality of the scan all contribute to reduced accuracy. Cloud-based OCR services like Google Cloud Vision have made improvements in this area by using deep learning, but they still generally require additional pre- and post-processing—for instance, denoising, normalisation, and correction—before the results can be used. More recently, multimodal large language models such as Gemini and those in the GPT family have shown that it is possible for them to carry out handwriting OCR without needing to rely on any prior examples and to do so with high accuracy, managing to read complex handwriting while at the same time preserving the document's structure and context—this is the reason why Gemini 2.5 Flash has been selected as the OCR engine for this framework.

C. Methods for prompting and grading using large language models
Since plain similarity scoring cannot reason or understand language, large language models are able to judge not only the surface similarity but also the logical structure and completeness of the responses. Methods such as Plan-and-Solve prompting divide the marking process into a number of reasoning stages in order to reduce the errors that occur with single-shot prompting, while "LLM-as-a-Judge" arrangements evaluate the answers based on clear criteria that cover conceptual understanding, structure, and clarity. More recent multi-agent approaches—in which several AI assessors grade independently before coming to a consensus—have shown greater agreement with human graders than a single model acting by itself.

D. The study of the document's layout and the assessment of the diagrams
Actual answer sheets are rarely just plain text — in addition to the handwriting, they also contain diagrams, equations, and tables, and ordinary OCR is unable to properly separate these various elements, leading to a reduction in the quality of the extraction. For this reason, Document Layout Analysis (DLA) has now developed into an important preprocessing step. Recent research, which combines YOLOv8 with Convolutional Block Attention Modules (CBAM) and Particle Swarm Optimization (PSO), has enhanced the detection of handwritten sections, diagrams, and corrections before recognition begins. Although graph-based methods, CNNs, and key point detection have been used in the scoring of diagrams specifically, combining an understanding of diagrams with text-based semantic grading remains an open issue, especially in the case of STEM subjects which involve a large number of figures and formulas.

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410051

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

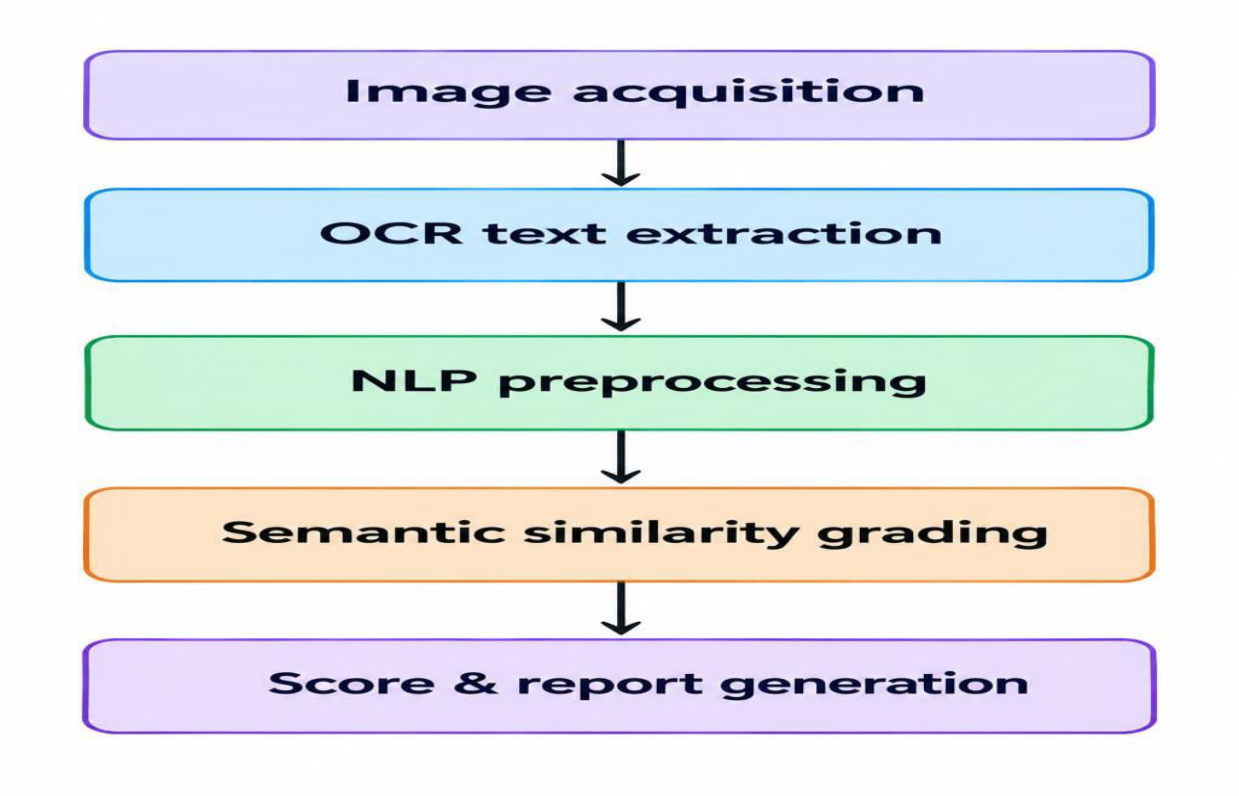
International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

E. Research Gaps

There are three disadvantages. While the majority of the current systems focus either on OCR accuracy or on semantic evaluation, they seldom address both of these aspects within a single, unified process. Diagrams, equations, and answers that contain a mix of content are still generally dealt with inadequately. Furthermore, most of the published systems have been tested only on small datasets and with only a limited comparison against real examiners, meaning that their practical reliability is still in doubt. The system described in this paper is meant to eliminate these weaknesses by combining multimodal OCR, NLP preprocessing, SBERT-based similarity assessment, rubric-based grading, and automated reporting into one pipeline which is designed to attain both consistency and scale.

III. PROPOSED SYSTEM ARCHITECTURE



The system that has been proposed includes five stages—image acquisition and preprocessing, the use of OCR and NLP preprocessing to extract the text, the grading of semantic similarity, and the production of the report—each stage being intended to enhance overall accuracy and consistency, and thus combining AI, OCR, and NLP into a single pipeline with the aim of achieving human-level reliability.

A. The acquisition and preprocessing of images

The process begins by using scanned or PDF versions of the answer sheets since the quality of the input affects all the subsequent steps; in order to reduce the computational demand without losing the details of the handwriting, the sheets are changed to grayscale, noise and any scan artifacts are eliminated, skew is corrected so that the lines of text appear horizontal, and the contrast and sharpening are carried out so that the OCR model can more easily read the characters. It is due to this combination of steps that a considerable reduction in recognition errors takes place later on.

B. OCR-Based Text Extraction

The text has been extracted by means of Gemini 2.5 Flash, a multimodal large language model that carries out zero-shot OCR by integrating the information from both the text and the visual elements rather than depending solely on character-recognition algorithms; it is for this reason that it is more reliable than Tesseract or standard cloud OCR services when dealing with different handwriting styles. The question numbers are automatically given to the extracted answers, and it is this assignment of question numbers that makes it possible to carry out question-by-question grading in the later stages.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. NLP Preprocessing

The text must first be cleaned if it is to be meaningfully compared; stray symbols, extra spaces, and OCR errors are removed, the capitalisation is made consistent and long answers are broken down into separate sentences so that the semantic comparison can be carried out more accurately. It is then put through lemmatisation, a process whereby words are reduced to their base forms without altering their meaning so that the system can identify the same concept even if it is expressed in different grammatical forms. The final outcome is clean text that is ready for the similarity stage.

D. Semantic Similarity Grading

At the heart of the system, both the student's answer and the model's answer are turned into dense vectors by means of SBERT, a method unlike keyword approaches in that it is able to capture the meaning and the relationships within entire sentences. The cosine similarity between these two vectors then produces a score indicating how conceptually close the student's answer is to the reference answer, rather than merely counting the number of words in common. Because of this capability, the system can assess conceptual accuracy, completeness, and the logical flow of the answer, not just whether certain keywords are present, and the score is later assigned to a rubric made up of categories such as Excellent, Good, Average, Fair, and Poor — leading to more consistent and fair results than are generally possible with manual grading.

E. The addition up of the scores and the compilation of the reports

When the questions have all been marked, the system calculates the total scores for the entire answer sheet and then produces a report automatically by means of a PDF library such as ReportLab or FPDF. The report contains the student's details, the answer extracted for each question, the similarity scores, the marks awarded, the total score, and an overall summary – thus providing instructors and students with a clear record of both what was extracted and how it was scored, which in turn aids future auditing and analysis.

Advantages of the Proposed Architecture

- It greatly reduces the time required for evaluation.
- It eliminates the human bias and inconsistency involved in grading.
- Provides an objective and repeatable means of assessment.

IV. EVALUATION CONSIDERATIONS

If the system is to be used in the real world, it first has to be validated using standard metrics and then be evaluated by human graders to ensure that it produces accurate and consistent results.

A. Performance Metrics

1) Precision

The extent to which the system considers answers to be relevant and whether or not they actually are relevant is called precision, meaning that there are fewer false positives when the answers are assessed.

2) Recall

The more relevant answers the system actually identifies—the more relevant answers it selects—the fewer correct ones will be missed.

3) F1-Score

A figure that is well balanced and particularly useful in the case of a skewed dataset can be achieved by calculating the harmonic mean of precision and recall.

4) Root Mean Square Error (RMSE)

The greater the similarity between the marks assigned by the AI and those provided by experienced human graders, the more accurate the system is at matching manual grading.

B. Human Evaluation

Instead of comparing the scores with those of a single grader, the scores given by the system should be compared with the average obtained from a number of experienced examiners, in order to minimise the impact of any one grader's personal tendencies and to establish a more reliable standard.

C. System Validation

The process of testing should include answer sheets drawn from a variety of subjects and featuring different handwriting styles, and it should examine the accuracy of the OCR, the similarity scores, the consistency of the grading, the processing time, and the degree of agreement with human graders—only after all these factors have been considered can the system be trusted for use in different academic situations.

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410051

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. CHALLENGES AND LIMITATIONS

A. Handwriting Variability

There is a great deal of variation between students in the way they write—whether it is in font size, spacing, or speed—and poor handwriting, overlapping strokes, or low-quality scans all cause the accuracy of OCR to decrease, which in turn affects all the subsequent steps.

B. The evaluation of the diagram and the mathematical expression

The STEM exams make extensive use of diagrams, graphs, and equations, and as the current semantic evaluation method is based on text it cannot interpret them, answers which are heavily dependent on diagrams still lie beyond the present capabilities of the system.

C. Limited Human Validation

Most AI grading systems, together with other similar ones that have been mentioned in the literature, have been tested only on small datasets with only a limited comparison to actual examiners — therefore, wider validation is required before they can be widely deployed.

D. Subjective Answer Assessment

It is difficult to give marks to answers that are only partly correct, even if they are put forward in an inventive way or even if the reasoning is implied — this limitation still applies even with powerful semantic models and LLMs.

E. Dataset Availability

Actually, there are very few large public datasets consisting of handwritten subjective answers that have annotations at the level of a rubric, and this makes it more difficult than it should be to train and benchmark such systems

VI. FUTURE SCOPE

The extraction process on more complicated sheets could be improved by employing attention-based YOLOv8 architectures in such a way that text, diagrams, equations, and corrections can be more accurately separated before OCR is carried out.

The process of grading could involve the use of a multi-agent large language model system, with a number of AI assessors looking at various aspects such as relevance, correctness, reasoning, and clarity before reaching a final score that was agreed upon by all of them — a method which has already been found to show greater agreement with expert graders than scoring done by a single model.

Other valuable features would be multilingual grading, rubrics that are automatically generated according to the type of question, adaptive scoring, personalized feedback, plagiarism checks, and cloud deployment for large university-scale exams – all of these features would make the system more flexible and suitable for large-scale use.

VII. CONCLUSION

To sum up, this paper has looked at the use of an AI model for the assessment of handwritten answer sheets, the model making use of multimodal OCR and SBERT to determine semantic similarity against a grading rubric based on conceptual proximity rather than on keyword matching. By combining transformer-based language models with multimodal OCR and incorporating semantic analysis, the model overcomes the issues associated with manual evaluation of written exams such as inconsistencies, evaluator fatigue, and long processing times. Although the model represents a solid starting point, there are still a number of challenges to deal with concerning handwriting variability, the evaluation of mathematical expressions, diagrams, and the need for extensive testing. Nevertheless, with progress being made in multimodal AI and semantic reasoning, automation in the evaluation of handwritten text is well on the way to becoming a viable reality.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

[1] A. Kulkarni, R. Patil, S. Endait, G. Kale, and R. Ghatage, "Automated Answer and Diagram Scoring in the STEM Domain: A Literature Review," in Proc. 2024 5th Int. Conf. Emerging Technology (INCET), Karnataka, India, May 2024, doi: 10.1109/INCET61516.2024.10593060.

[2] B. Aarathisree, S. Sarkar, B. D. Dammala, M. Panday, and N. Sharma, "Revolutionizing the Future of Automated Subjective Answer Sheet Evaluation System with Machine Learning and LLMs," in Proc. 2024 IEEE Pune Section Int. Conf. (PuneCon), Pune, India, Dec. 2024, doi: 10.1109/PuneCon63413.2024.10895748.

[3] Palak, Mohit, N. Sharma, Bhushan, and S. Jain, "Python-based OMR Sheet Recognition and Evaluation," in Proc. Int. Conf. Modern Sustainable Systems (CMSS), 2025, doi: 10.1109/CMSS66566.2025.11182432.

[4] N. A. F. Fansab, S. A. I. Alfarozi, and I. Ardiyanto, "Improvement of Small Object Detection of Handwritten Answer Sheet," in Proc. 2025 Int. Conf. Advancement in Data Science, E-learning and Information System (ICADEIS), 2025, doi: 10.1109/ICADEIS65852.2025.10933354.

[5] S. Sarkar, R. Jayadevan, N. Sharma, and M. Pandey, "Enhancing Handwritten Answer Sheet Evaluation Through Large Language Models and Multi-Agent Reasoning," in Proc. 2025 IEEE Pune Section Int. Conf. (PuneCon), Pune, India, Dec. 2025, doi: 10.1109/PUNECON67554.2025.11379082.



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details