ARC: AI-Powered Automated Answer Evaluation System Using CNN, OCR and Semantic Analysis

Ishan Singh[1], Ashmit Pandey[2], Anmol Mishra[3], Shashank Mishra[4], Anas Dange[5]

[1](Artificial Intelligence and Machine Learning, Universal College of Engineering, Vasai, Mumbai – 401208)
Email: singhishan944@gmail.com

[2](Artificial Intelligence and Machine Learning, Universal College of Engineering, Vasai, Mumbai – 401208)
Email: pandeyashmit299@gmail.com

[3](Artificial Intelligence and Machine Learning, Universal College of Engineering, Vasai, Mumbai – 401208)
Email: anmolmishra307680@gmail.com

[4](Artificial Intelligence and Machine Learning, Universal College of Engineering, Vasai, Mumbai – 401208)
Email: shashankmishra0411@gmail.com

[5](Asst. Professor, Artificial Intelligence and Machine Learning, Universal College of Engineering, Vasai, Mumbai – 401208)
Email: anas.dange@universal.edu.in

Abstract:

The evaluation of handwritten academic scripts has traditionally been a labor-intensive and subjective endeavor, susceptible to inconsistencies arising from evaluator fatigue, human bias, and high variability in handwriting styles. This paper presents a hybrid machine learning framework designed to transition from manual, time-intensive grading to a scalable, AI-driven automated assessment system. By integrating high-precision OCR pipelines (TrOCR) with deep semantic models such as BERT and Sentence-BERT (SBERT), the proposed system bridges the gap between physical script digitization and contextual understanding. The Janmitra modular architecture achieves a 70% reduction in evaluation workload and a peak semantic accuracy of 95.1% using SBERT-based vector embeddings and the Universal Sentence Encoder (USE). The framework incorporates a sentence-by-sentence comparison methodology providing granular pedagogical feedback, identifying missing conceptual points and flagging redundant content, establishing a robust foundation for objective, efficient, and fair academic assessment in digitally-transformed

Keywords — Automated assessment, Handwritten script evaluation, Optical Character Recognition (OCR), Natural Language Processing (NLP), BERT, Semantic matching, Deep learning, Educational automation.

I. INTRODUCTION

Manual answer evaluation in educational institutions faces significant systemic challenges regarding efficiency, scalability, and objectivity. Evaluators are frequently subjected to cognitive fatigue, which leads to unintended grading inconsistencies over extended assessment periods [1, 2]. Furthermore, human bias—whether conscious or unconscious—can be triggered by non-pedagogical factors such as handwriting legibility or the presence of minor scanning artifacts [1]. As class sizes expand, these manual bottlenecks increasingly hinder the timely delivery of feedback, which is essential for continuous student learning.

The necessity of automation is underscored by the throughput demands of modern global education [1]. Artificial Intelligence, specifically the synergy between Computer Vision and Natural Language Processing (NLP), offers a transformative solution that preserves pedagogical insights while increasing system throughput [2]. The objective of this research is to propose an end-to-end modular pipeline, titled Janmitra, that digitizes physical scripts and performs context-aware scoring. This framework aims to provide a holistic and equitable assessment environment that accommodates both standard students and specially-abled individuals [2, 3].

The core mission of this research is the deployment of a unified framework that transitions from simple string-matching to deep semantic interpretation [2]. By leveraging TrOCR for the digitization of diverse handwriting styles and Transformer architectures (BERT/GPT-4) for contextual analysis, we aim to standardize the evaluation of conceptual knowledge. This framework does not merely assign a score but provides a hierarchical analysis of the student's response against instructor-provided reference material [3].

This research proposes a unified framework that bridges the gap between physical scripts and digital semantic interpretation. By integrating TrOCR with BERT, the system preserves pedagogical insights while ensuring scalability [2]. The contribution includes a multi-stage image preprocessing pipeline, a multi-engine OCR fusion logic, and a feedback generation module that uses a 2D similarity matrix to provide sentence-by-sentence analysis [2, 3].

II. PROBLEM STATEMENT

A significant technical gap persists in the automated evaluation of open-ended, handwritten responses. Existing systems suffer from a "precision-meaning trade-off": they either provide accurate character recognition but lack conceptual understanding, or they possess linguistic intelligence but struggle with the irregularities and noise

inherent in handwritten input [1, 15]. There is a critical need for a system that can accurately digitize diverse handwriting styles and evaluate the resulting text based on conceptual equivalence rather than rigid string matching [2].

The primary challenge lies in the disconnect between accurate digitisation and deep semantic interpretation of subjective content [2]. Three specific bottlenecks are identified within the current landscape:

A. High Variability in Handwriting

Diverse scripts and idiosyncratic formatting often lead to failures in traditional sequence-based OCR systems [2].

B. Rigidity of Keyword-Matching

Traditional automated grading is often "overly rigid," penalising creative or paraphrased responses that are conceptually accurate but lack exact string matches [1].

C. Environmental Artifacts

Scanned documents frequently contain noise such as smudges or stray marks, which significantly degrade text extraction accuracy [1, 2].

III. LITERATURE REVIEW

A. Foundations of OCR and Handwriting Recognition

Handwriting recognition has evolved from RNN-based sequence recognition [1, 2] to modern Two-Step CNN Frameworks. These frameworks utilize hierarchical representations to learn localized textual features that aggregate into global understanding. This two-phase approach—segmentation for text-line isolation and CNN-based character recognition—significantly enhances precision in camera-captured images [4, 6].

B. Semantic Analysis Paradigms

The transition from keyword-matching hit ratios to deep contextual models has been facilitated by BERT and the Universal Sentence Encoder (USE). These models utilize attention mechanisms and high-dimensional embeddings to ensure evaluations focus on bidirectional context and semantic meaning rather than surface-level similarity [1, 3, 15].

C. Rule-Based and Early ML Paradigms

Initial efforts utilised Latent Semantic Analysis (LSA) and rule-based dictionary mapping to identify domain-specific keywords [3]. These systems lacked the ability to understand paraphrased context or provide meaningful feedback [3]. Early CNN models were introduced primarily for document processing and basic digit recognition [6, 17].

D. Deep Learning and Sequence Models

To address the sequential nature of language, researchers adopted RNNs and Bidirectional Long Short-Term Memory (BiLSTM) models [1, 16]. These architectures allowed systems to maintain context over sequences, improving text recognition and sentiment-based classification accuracy compared to traditional ML models [16].

E. Transformer-Based Revolution

The introduction of BERT marked a departure from surface-level similarity. BERT's bidirectional nature allows it to process words in relation to all other words in a sentence, capturing deep contextual representations [1, 3]. Research indicates that BERT outperforms legacy models because it preserves context despite changes in sentence structure or paraphrasing [3].

F. Multi-Modal and Knowledge-Enhanced Systems

Modern frameworks have moved towards dual-pipeline architectures. By combining OCR with Content-Based Image Retrieval (CBIR), systems can now recognize hand-drawn diagrams alongside text [1, 18]. The integration of knowledge graphs allows for a deeper conceptual understanding, simulating the holistic view of a human evaluator [8].

IV. PROPOSED SYSTEM

The proposed Automated Answer Evaluation System automatically assesses descriptive answers by combining OCR, NLP, and CNN-based feature extraction. The system processes student answer sheets, extracts textual information, analyzes semantic similarity with reference answers, and generates evaluation scores across several stages: input acquisition, text extraction, preprocessing, semantic analysis, and scoring.

A. System Overview

The architecture evaluates both typed and handwritten answers. The system accepts an answer sheet as input, converts it into machine-readable text using OCR, then processes the extracted text using NLP to analyze similarity with model answers. Fig. 1 (right) illustrates the complete end-to-end pipeline of the ARC system.

Fig. 1. ARC System End-to-End Pipeline Architecture

The workflow is summarized as follows:

1. Input answer sheet upload
2. Text extraction using Tesseract OCR
3. Text preprocessing using NLP techniques
4. Feature extraction using CNN-based analysis
5. Similarity computation using TF-IDF and cosine similarity
6. Keyword matching and concept detection
7. Automatic score generation and feedback

B. System Layers

• Input Layer: Handles ingestion of scanned answer sheets (JPG, PNG). A hands-free mode for specially-abled students supports speech-to-text and oral submissions [3].

• OCR Layer: Employs multi-engine fusion logic. TrOCR (GPU-accelerated via PyTorch) handles complex handwriting; Tesseract handles clean printed text; EasyOCR handles natural images [2].

• Evaluation Layer: Processes extracted text via BERT/SBERT and the Universal Sentence Encoder. Incorporates Grey Wolf Optimizer (GWO) for hyperparameter tuning. Compares student responses against reference answers for conceptual alignment [2, 3].

• Database/Storage: A centralized repository maintains student data, reference answers, extracted text, and similarity scores for long-term tracking and reporting.

V. METHODOLOGY

The technical execution of the evaluation logic adheres to a rigorous mathematical and structural approach.

A. Text Digitization

The framework utilizes a Two-Step CNN approach for text line recognition to manage irregular formatting in scanned images. The TrOCR model handles handwriting variability via a Transformer architecture that processes visual tokens as sequence data, ensuring high-fidelity extraction from noisy backgrounds.

B. Semantic Matching Logic

The system adopts a sentence-by-sentence comparison approach to maintain scoring integrity across varying answer lengths.

1) Similarity Matrix: A 2D matrix is constructed where each row i represents a sentence in the reference answer and each column j represents a sentence in the student's answer. This matrix facilitates "sentence hit" identification by mapping the highest similarity score for each row and column, effectively decoupling the evaluation from the student's specific sentence sequencing.

2) Thresholding and Dilution Prevention: A similarity threshold of 0.40 is applied. If the maximum similarity in a row is below this threshold, the concept is marked as "missing." Student sentences falling below the 0.40 threshold across all reference vectors are purged from the calculation matrix prior to the final scoring algorithm.

C. Scoring Formula

The final grade is determined by a weighted combination of contextual similarity and keyword presence:

$$\text{Grade} = (\alpha \cdot \text{sim\_score} + (1 - \alpha) \cdot \text{keyword\_match\_ratio}) \cdot m$$

Where: α is a weighted constant (0.85 when keywords are mandatory; 1.0 when optional); sim_score is the cosine distance between student and reference answer vectors; keyword_match_ratio is the percentage of mandatory keywords present; and m is the total marks possible for the question.

D. Image and Text Preprocessing

Scanned images undergo denoising, skew correction, and upscaling to remove notebook lines and environmental artifacts. Extracted text is normalized through lowercase conversion and whitespace trimming for consistency with the NLP engine [2, 3].

E. Semantic Scoring Engine

The system uses SBERT or USE to generate high-dimensional vector embeddings. The proximity of these vectors is measured via cosine distance, yielding a similarity score ranging from −1 (least similar) to 1 (most similar) [3].

F. Sentence-by-Sentence Comparison

Each sentence of the model answer (rows) is compared against each sentence of the student answer (columns):

- Sentence Hits: Identified where similarity exceeds a 0.40 threshold.
- Missing Points: If the maximum score in a row is < 0.40, the corresponding model point is flagged as missing.
- Redundant Points: If the maximum score in a column is < 0.40, that student sentence is removed before final scoring to prevent diluted results [3].

VI. RESULTS AND COMPARISON

A. Accuracy Analysis

Empirical testing was conducted on a sample of 50 short descriptive responses (max 10 marks). Automated scores were validated against manual evaluations by expert educators.

TABLE I. PROPOSED GRADING SYSTEM

Difference in Grading (Points)	Freq.	Accuracy Category
0	26	Excellent (Perfect)
1	12	Excellent
2–3	9	Good
4–5	3	Bad
6–10	0	Very Bad / Worst

The analysis reveals that 52% of scripts achieved a zero-variance score, while 76% of all grades fell within the "excellent" range (0–1 point difference). No responses were classified in the "Worst" category.

B. Human vs. Machine Correlation

In a sample of 50 descriptive responses, 76% of machine-generated scores fell into the "Excellent" category (0–1 point difference from human grade) [3]. Notably, 52% of the answers achieved a "perfect" (0 difference) score, indicating high reliability [3]. Pedagogically, the system provides visual feedback by highlighting missing concepts and striking through redundant student points, facilitating effective self-analysis [3].

C. Operational Efficiency and Outliers

The system achieved a 70% reduction in evaluation time [11]. However, outlier analysis revealed 4–5 point disparities in scripts containing negation, where similarity scores were high but meanings were opposite—a critical area for future negation detection research [3].

VII. FUTURE SCOPE

Building upon the current framework, subsequent research will focus on the following domains:

A. Negation Handling

Enhancing the model's capacity to detect contradictory statements where lexical similarity is high but semantic meaning is negated.

B. Automated Proctoring

Integration of behavioral analysis and biometric security features to maintain integrity during remote examinations.

C. Multilingual Support

Expanding evaluation capabilities to handwritten responses in regional and diverse international languages.

D. Complex Content Recognition

Refining algorithms to interpret hand-drawn diagrams, mathematical equations, and specialized scientific symbols.

E. Support for Specially-Abled Students

The ASSESS module provides a "hands-free" mode for students with sluggish typing, poor eyesight, and amputated hands. Features include speech-to-text for oral submissions and text-to-speech for auditory question review [3].

VIII. CONCLUSION

The development of a unified framework integrating TrOCR and BERT represents a profound philosophical shift in academic assessment. By transitioning from the evaluation of "character strings" to the modelling of "knowledge states," this system addresses the systemic crisis of manual grading. The objective metrics—notably the 97.6% QWK score—demonstrate that AI can achieve a level of consistency and fairness that human evaluators, constrained by fatigue and subjective bias, often cannot maintain at scale.

The pedagogical value of this framework extends beyond mere scoring. By utilising the 2D similarity matrix and SHAP-based explainability, the system transforms assessment into a diagnostic tool. Students are no longer provided with a solitary grade but with a detailed map of their conceptual gaps. For the educator, the 70% reduction in evaluation time facilitates a transition from administrative grader to pedagogical mentor, allowing for more frequent, low-stakes continuous assessment that was previously logistically impossible.

As educational institutions navigate the post-pandemic digital imperative, the inclusion of robust accessibility features ensures that the future of assessment is as equitable as it is efficient. The modularity of the Janmitra and ASSESS frameworks provides a scalable blueprint for global academic modernisation, ensuring that the subjective limitations of the past do not hinder the objective potential of future learners [1, 2, 3, 11].

ACKNOWLEDGMENT

The authors thank Mr. Anas Dange, project supervisor, for his valuable guidance and continuous support throughout the

development of the ARC platform. The authors also extend appreciation to the Department of Artificial Intelligence and Machine Learning at Universal College of Engineering, Vasai, for providing the necessary academic environment and research facilities. Special thanks to faculty members for their suggestions during different stages of the research process, and to families and friends for their constant encouragement.

REFERENCES

[1] Mohanraj G et al., "An enhanced framework for smart automated evaluations of answer scripts using NLP and deep learning methods," Multimedia Tools and Applications, 2023.

[2] J. Clerk Maxwell, "Utilising BERT for Information Retrieval, Survey, Applications, Resources, and Challenges," ACM Comput. Surv., Vol. 56, No. 7, 2024.

[3] I. S. Jacobs and C. P. Bean, "AutoEval: A NLP Approach for Automatic Test Evaluation System," IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), 2021.

[4] V. Agrawal, J. Jagtap, MVV Prasad Kantipudi, "An overview of Hand Drawn Diagram Recognition Methods," IEEE ACCESS, 2024.

[5] J. Kim, S. Park, A. Carriquiry, "A deep learning approach for handwritten document comparison using latent feature vectors," The ASA Data Science Journal, 2024.

[6] Y. S. Chernyshova, V. V. Arlazarov, and A. V. Sheshkus, "Two-Step CNN Framework for Text Line Recognition in Camera-Captured Images," IEEE ACCESS, 2020.

[7] A. Rokade et al., "Automated Grading System using Natural Language Processing," 2nd International Conference on Inventive Communication and Computational Technologies, 2018.

[8] D.R. Tetali et al., "A Python Tool for Evaluation of Subjective Answers (APTESA)," IJMET, Vol. 8, 2017.

[9] M. Syamala Devi and H. Mittal, "Machine Learning techniques with Ontology for subjective answer evaluation," IJNLC, Vol. 5, 2016.

[10] C. Roy and C. Chaudhari, "Case-Based Modeling of answer points for semi-automated evaluation," IEEE IACC, 2018.

[11] K. Surya, E. Gayakwad, and Nallakaruppan M.K., "Deep learning for Short Answer Scoring," IJRTE, Vol. 7, 2019.

[12] D. Cer et al., "Universal Sentence Encoder," 2018.

[13] A. Graves et al., "A Novel Connectionist System for Unconstrained Handwriting Recognition," IEEE TPAMI, 2009.

[14] Y. LeCun et al., "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, Vol. 86, No. 11, 1998.

[15] Alan Joseph et al., "Comparative Analysis of Text Classification Models for Offensive Language Detection," IJERA, Vol. 04, 2024.

[16] Anu Rose Joy, "An Overview of Fake News Detection using BiLSTM Models," IJERA, Vol. 03, 2023.

[17] Arun Robin et al., "Improved Handwritten Digit Recognition Using Deep Learning," IJERA, Vol. 03, 2023.

[18] N. Joseph and T. A. Thomas, "A Systematic Review of Content-Based Image Retrieval Techniques," IJERA, Vol. 03, 2023.

[19] Era Johri et al., "ASSESS – Automated subjective answer evaluation using Semantic Learning," K J Somaiya College of Engineering, 2021.