MUSE_E: Multimodal AI Framework for Automated Answer Script Evaluation: Integrating Handwriting Recognition and Semantic Scoring

[1]KAMALA V
ASSISTANT PROFESSOR,
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING. KGiSL INSTITUTE OF TECHNOLOGY, COIMBATORE, TAMIL NADU, INDIA
kamala.v@kgkite.ac.in
[2]MITRA K
UG STUDENT,
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING. KGiSL INSTITUTE OF TECHNOLOGY. COIMBATORE, TAMIL NADU, INDIA
kmitra17804@gmail.com
[3]PRADEEPA S
UG STUDENT,
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING. KGISL INSTITUTE OF TECHNOLOGY, COIMBATORE, TAMIL NADU, INDIA
pradeepashanmugasundaram04@gmail.com
[4]SUBA H
UG STUDENT,
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING. KGISL INSTITUTE OF TECHNOLOGY,
COIMBATORE, TAMIL NADU, INDIA
subaharichandran123@gmail.com
[5]SUREKA R
UG STUDENT, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING. KGISL INSTITUTE OF TECHNOLOGY.
COIMBATORE, TAMIL NADU, INDIA
surekar1204@gmail.com

ABSTRACT

The advancement of technologies such as Artificial Intelligence (AI) and Vision-Language Models (VLM) has significantly transformed academic assessments into an Automated grading system. This paper states about an AI- Powered Automated Answer script evaluation system that integrates secure web-based exam management, and it assists multimodal AI-driven grading using Flask as backend and React as frontend with JWT-based role authentication for teachers and students. In this application, teachers can create examinations with a pre-defined answer key, total marks, and submission timelines. On the other side, students can upload handwritten answer scripts as pdf or images. The evaluation engine utilizes the Groq LLaMA Vision Model to perform Optical Character Recognition (OCR) and semantic grading by extracting actual text content from encoded handwritten scripts and comparing it with the original answer key. The application evaluates the presence of keywords, concepts, diagrams, and semantic alignment to generate scores based on the answer key, along with detailed AI-driven feedback. It stores the details of marks in a database for retrieval and transparency. By integrating automated OCR, semantic scoring, and secure workflow management, the proposed framework strengthens the efficiency of grading, maintains consistency, scalability, and reliability, and notably reduces manual effort while ensuring fairness in digital academic assessment.

KEYWORDS – Artificial Intelligence (AI), Automated Answer Evaluation, Optical Character Recognition (OCR), Semantic Grading, Vision-Language Model (VLM)

I. INTRODUCTION

The scaling up of the students increases year by year , thus creates an atmosphere to increase the examination to test and improve them. As human evaluations of answer scripts takes lots of time and cause manual errors and evaluation varies from examiner to examiner and can be biased also. Because of these issues, the academic institutions are moving towards AI-MUSE_E to ensure fairer, faster and more consistent grading [4][9][10]. The early stages of automation in exams mainly used Optical Mark Recognition (OMR) which only works for multiple choice questions and the limitation of OMR is that it cannot evaluate descriptive or short answer questions [3][5]. The development in OCR systems still faces challenges due to handwriting variability and poor image quality to overcome these the researchers have developed techniques such as Confidence using CNNs, OCR error rejection, Post correction techniques, using Large Language Models (LLMs) to extract the text [8][12][14].For the precise evaluation of descriptive answer requires understanding the concept rather than keyword matching. The modern systems use semantic similarity techniques using word embeddings and transformer models like (Bidirectional Encoder Representations from Transformers) BERT and Hybrid Natural Language Processing (NLP) to assess answer quality [1][6][13]. This approach helps in better understanding of students answer scripts and improve grading reliability with co-relation with human examiners [5][13]. The recent research highlights the multimodal fusion in integrating the image based handwritten recognition with language understanding for complex evaluation tasks [4][8]. Despite the fact that existing automated answer evaluation applications gives in a promising results [1][9][10]. The systematic literature reviews indicate consistent challenges related to generalization, robustness and real-world deployment in academic institutions [2][11]. To overcome existing limitations, the paper proposes MUSE_E by integrating handwritten recognition and semantic scoring within a cohesive architecture. MUSE_E aims to enhance evaluation accuracy, robustness, and scalability for handwritten descriptive answer assessments.

II. LITERATURE REVIEW

A. Comprehensive Analysis of Research Papers

The Comprehensive analysis of 14 research papers is outlined in the table below. This interpretation focuses on key areas such as automated answer script evaluation, HTR, OCR, and AI-assisted semantic scoring in educational assessment systems. This study addresses algorithms or techniques employed, the primary challenges, the significant features, and the specific research gaps highlighted.

TABLE I. Literature Review Summary of Related works

S.NO	TITLE	YEAR	METHODOLOGY	GAP(from conclusion/future work)
1.	Automated Answer Sheet Evaluation Using OCR and NLP.	2025	Uses EasyOCR for text extraction and BERT-based NLP for semantic analysis with cosine similarity using Flask and React+JS.	The model performance depends on OCR accuracy and has limitations towards short descriptive answers, not essays or diagrams. Future work addresses over using large datasets for semantic understanding.

2.	Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR).	2020	Systematic Literature Review was conducted using Kitchenham protocol and analyzed 176 selected studies focusing on OCR methods on handwritten, datasets and languages.	It shows gap in multilanguage OCR, dataset diversity, and Deep Learning(DL). So it suggested the research on hybrid AI-models and language independent OCR technology.
3.	Recognition of Machine - Readable Zone in Identity Documents: A Review.	2025	Survey of (Machine Readable Zone) MRZ recognition techniques, MRZ-specific postprocessing and error correction.	Require for MRZ specialized recognition and multi-language models, and handling photos with projective blurs; computationally-light on- device solutions. The model to be covered on the areas of MRZ dataset targeted error correction modules.
4.	Recent Advances and Trends in Multimodal Deep Learning: A Review.	2021	This multimodal is an architecture for fusion, alignment, representation;covers image/text/audio/video/physio; reviews datasets, and evaluation metrics.	The gaps in this model is data efficiency in multimodal E2E models, and scarcity of multimodal datasets. Future improvement on frontend linguistic features and novel datasets.
5.	Automated S hortAnswer Grading using Semantic Similarity based on Word Embedding (IJTech).	2021	Methods proposed on semantic similarity via word embeddings + syntactic analysis to compare student answer with the reference answer key.	Gaps requires different reference answers and limited handling of varied phrasing and domain specific terms. Future requirements, it needs richer semantic models, larger multireference approaches towards domain.
6.	Automated grading using natural language processing and semantic analysis (MethodsX).	2025	This model of hybrid automated answer evaluation system that combines lexical similarity, normalized word-count, and deep semantic similarity using TensorFlow Universal Sentence Encoder. This system calculates waited final score and applied rule- based thresholds to assign marks.	The limitations of this model, it tested on very small number of answer, so the results are not fully accurate, and it needs to be tested on large dataset for accuracy. Using bigger and more diverse dataset help in confident scoring, and integrating this system with OCR module helps in scannedhandwritten answer sheets.
7.	Handwritten Text Recognition: A Survey	2025	The research has been done on HTR systems. These systems were grouped based on their recognition levels (word level, line level, paragraph level, document level). The study reviewed on different AI technology such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Transformer architectures and benchmarking datasets.	This system requires large amount of labelled data for proper training and need improvement in layout understanding, multilingual adaptation, and end-to-end models.
8.	uFOIL: An Unsupervised Fusion of Image Processing and Language Understanding	2025	This system uses Unsupervised ensemble pipeline, the process includes image processing, ensemble of OCR systems, confidence scoring, and Generative Adversarial Network (GAN) augmentation for data variability.	This system fails to handle unstructured documents such as handwritten answer scripts. It does not support Multilanguage properly. It is not trained using different datasets. Future enhancement is to recognize freeform handwriting, optimize the system to work faster.
9.	Review on Smart Evaluation of Descriptive Answer Sheets	2024	This proposed system uses AI, Machine Learning (ML), and NLP to automatically evaluated long answers. This system uses existing models like BERT, CNN-LSTM	The current system lacks strong semantic understanding and not good for real-world adaptability. So, this system aims improvement in NLP-based
			(Long Short-Term Memory) model, GPT-based summarization models.	scoring accuracy and to ensure data privacy and security as future enhancement.
10.	A Systematic Literature Review: Are Automated Essay Scoring Systems Competent in Real - Life Education Scenarios?	2024	Systematic Literature Review (SLR) was conducted and reviewed over 104 research papers related to Automated Essay Scoring (AES). The researches followed Keele SLR protocol that use inclusion/exclusion criteria and quality assessment checklist.	The current AES system are not fully capable of scalability, accuracy, and robustness. The future improvement focused on enhancing feedback generation for students, improving reliability, and increasing domain adaptability.
11.	Systematic Literature Review of Machine Learning Models and Applications for Text Recognition.	2025	This system presents a taxonomy and time-based analysis of OCR models and tracing them from the traditional evolution of ML methods to CNN/LSTM architectures → GAN based approach and finally transformer-based models.	The limitations of this system includes limited availability of high quality, multi-script datasets and supports less languages and this model overfit due to small training datasets. Additionally it can't support real time deployment as well as errors caused by character segmentation issues. To compress these limitations the future research will focus on self- supervised pretraining, and development of TinyML based light-weight models, and multimodal OCR systems combined with robust post-processing pipelines.

12.	The debate on automated essay grading (Trends & Controversies — overview / commentary).	2020	This work presents scholarly review combined with commentary from multiple authors, which offers an overview of automated essay evaluation system such as PEG, e- rater, and LSA. It explains automatic feature extraction methods, distinguishing between proxy features and direct linguistic features. The review also discusses about the psychometric validity and also examines their educational and pedagogical implications.	This paper overcomes many limitations of current automated evaluation systems that include restricted ability to accurately assess rhetorical structure, essay organization, and higher- level content quality. In addition, the feedback provided is often shallow and non-diagnostic, which provides guidance for meaningful improvement. The future enhancement will be based on delivering localized, diagnostic feedback, demonstrate awareness of rhetorical and structural elements.
13.	A Hybrid Approach for Automated Short Answer Grading	2024	This study presents a hybrid BERT based Automatic Short Answer Grading (ASAG) model that integrates custom multi-head attention, parallel CNN layers, and	Future research will focus on improving computational efficiency, evaluating the model on large multilingual datasets,
			LSTM network to enhance context understanding and grading accuracy. The model's effectiveness is validated through experiments on standard benchmark datasets, provides an improved grading performance.	And incorporating explainable AI techniques to make grading decision transparent.
14.	Review on Smart Evaluation of Descriptive Answer Sheets	2024	This work presents a review on system that applies AI, ML, NLP to automate the evaluation of descriptive answers. It compares existing approaches based on CNN- LSTM, BERT, GPT-based models for summarisation techniques to analyse their effectiveness in accessing student responses.	The limitations of these system is it has limited and non-comprehensive datasets, insufficient semantic understanding of student answer and poor adaptability in real-time evaluation scenarios. The future enhancement focuses on enhancing NLP-based scoring accuracy while also ensuring data privacy and security in educational AI systems.

From the above table, it can be inferred that existing automation evaluation of answer scripts is improving quickly by using technologies such as OCR, DL, NLP, and multimodal AI. Even though there are many improvements, it still has many practical problems in real world. The present systems struggle to handle different languages, varied handwriting styles, and OCR errors and they fail to understand the meaning of the complex, creative, a lengthy answers. Most of the models depends upon predefined reference answers, which makes fair assessment difficult in real classroom situations. Additionally, it concerns about the data privacy, bias, transparency, and high computational cost limit their wide spread use in educational institutions. Overall, while technical progress is promising, these systems are not yet robust or pedagogically effective enough. Hence, future research aims to build hybrid multimodal and use of large and diverse datasets improving the conversion of handwriting into meaningful semantic representation, increase explainability, and develop scalable grading systems that can also provide useful feedback to students.

III. MUSE_E: MULTIMODAL AI FRAMEWORK FOR AUTOMATED ANSWER SCRIPT EVALUATION: INTEGRATING HANDWRITING RECOGNITION AND SEMANTIC SCORING

A. System Overview

MUSE_E is an automated system developed to evaluate student written answer script efficiently and fairly. Instead of relying on slow and biased manual correction, the system uses OCR to convert handwritten or printed answers into digital text and then apply intelligent text comparison method to access student response against predefined answer key.

- Implementation: This web-based application uses Flask framework in python, MUSE_E allows professors/staff to upload question papers, answer key, and upload student answer scripts in organized manner. MUSE_E automatically process these inputs, evaluate answer, calculate course, and stores the result in centralized database.

Overall the purpose of MUSE_E is to save time, reduce human error and bias, to ensure consistent grading for handwritten records and maintain well organized academic records.

B. System Architecture

The architecture of MUSE_E follows a structured design ensuring different function oriented content and improve maintainability. Each function modules perform as specific function while interacting cohesively with other module to complete the evaluation process.

i) Application Layer: The application layer is developed Flask framework which serves as a central control unit of the system, handling routing, request processing and execution of evaluation logic. The main application file synchronize interaction between file upload, OCR processing, database processing, and stores computation. This structured backend helps in proper implementation of communication between component and efficient request handling.

ii) Database Layer: The database layer is implemented using SQLite which provides an effective relational database. This database helps in storing the information in structured way such as question paper, answer key, student details, text, and evaluation results. Database model defines relationship between entity ensuring proper connection between student submission and corresponding scores. This hierarchal structure helps in easy retrieval of evaluation records.

iii) OCR processing model: The OCR module plays an important role in converting the scanned written answer scripts into a machine-readable text. The uploaded images or pdf are processed through the OCR engine, which extracts the textual content from both the printed and the written answer scripts. Earlier for recognition of text, image preprocessing techniques such as grayscale conversion, noise removal, and contrast enhancement are applied for the improving the accuracy of the system. The extracted texts are cleaned and formatted to remove the inconsistencies before being forwarded to the evaluation engine.

iv) File Management System: To maintain a proper hierarchy, the system uses a structured directories for storing uploaded documents. Question papers, student scripts, and processed submission are stored in separate folders. This organization ensures systematic file management, easy retrieval and prevents data overlap between different documents.

v) Working Mechanism: MUSE_E executes through a sequential workflow that ensures accurate and efficient evaluation of answer scripts with confidence scoring.

vi) Upload Phase: The process begins when the instructors uploads the question papers with the time constrain and the corresponding answer keys through the web interface. Students uploads their scanned written answer sheets in the supported formats with in the time limit. Once it is uploaded, the files are securely stored min the designated directories.

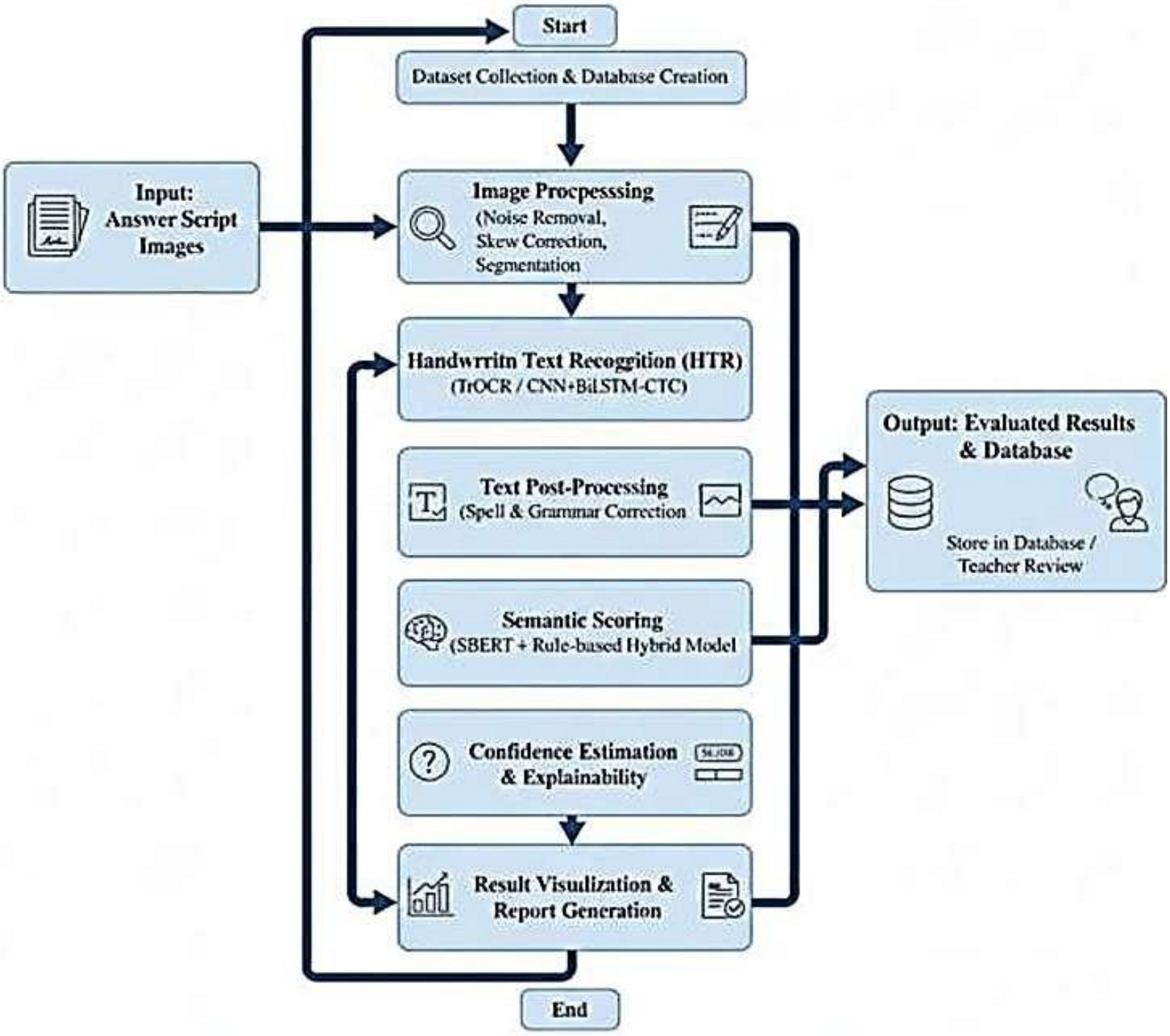
vii) Text Extraction Phase: After upload, the student's answer script is passed into the OCR module. The module processes the file and applies preprocessing techniques, and extracts textual data. The recognized text is then structured and formatted for analysis. The extracted text is divided based on the question numbering patterns and then checked with the answer key for accurate evaluation.

viii) Answer Comparison and Evaluation: The evaluation engine compares student answer script with the predefined answer key. The engine uses text normalization techniques such as tokenization and case standardization are applied before comparison. The system identifies matching keywords and relevant textual patterns between student's answer scripts and the original answer key. This method ensures fairness by awarding partial credit when appropriate.

ix) Result Generation and Storage: After evaluating all questions, the system calculates the total score and stores it in the database. The results are linked to the student portal for their review. The structured storage enables long-term maintenance of academic records.

x) Technical Implementation: The MUSE_E is developed using python because it offers strong library and tools for text processing and ML tasks. And it also offers lightweight, flexible, and well suited for handling the request efficiently in backend flask framework. For data storage, SQLite is used as it provides simple yet effective database. The system architecture is organized so that the main application file controls request routing and evaluation logic, while separate modules handle database operation and OCR functionality. Furthermore, a virtual environment is used to handle dependencies, which help maintain consistency when the project runs on different machines.

xi) Evaluation Logic: The evaluating process in MUSE_E is based on structured text comparison techniques. The system identifies essential keywords and concepts of student's answer scripts with the original answer key for scoring. The similarity score is calculated as a ratio between matched keywords and the total expected keywords, multiplied by the maximum marks allocated for the question. This proportional scoring approach enables partial marking and maintains consistency across evaluations. Although the system primarily relies on keyword matching, the framework can support more advanced similarity algorithms if integrated in future iterations.Security and Data Handling: MUSE_E ensures structured handling of uploaded records. Files are stored within controlled directories, and database entries maintain proper connections between students and their submission. The architecture allows integration of authentication mechanisms to restrict unauthorized access and safeguard academic data.


Start
Dataset Collection & Database Creation
Input: Answer Script Images
Image Procpessing (Noise Removal, Skew Correction, Segmentation
Handwrritn Text Recoggition (HTR) (TrOCR / CNN+BiLSTM-CTC)
Text Post-Processing (Spell & Grammar Correction
Semantic Scoring (SBERT + Rule-based Hybrid Model
Confidence Estimation & Explainability
Result Visulization & Report Generation
Output: Evaluated Results & Database
Store in Database / Teacher Review
End

IV. CONCLUSION & FUTURE ENHANCEMENT

The MUSE_E system successfully automates the evaluation of student answer scripts by combining OCR with structured text comparison methods. This automation greatly reduces the effort and time required for manual evaluation while ensuring consistent, unbiased, and well-organized assessment. Its modules design and database-supported workflow make the system reliable and suitable for digitizing academic evaluation processes in educational institutions. Despite some limitations in handling highly descriptive answers and unclear handwriting, MUSE_E provides a practical and scalable solution with intelligent automated assessment in modern education.

The future enhancements can include integrating deep learning–based handwriting recognition models to better process complex and varied scripts. Deploying the system on cloud platforms would allow large-scale institutional adoption. Additionally, plagiarism detection, and Learning Management System (LMS) integration can transform MUSE_E into a comprehensive digital examination and evaluation management solution.

V. REFERENCES

[1] B. Santhosh, A. Sagar, B. Manikanta, A. Abhiram, and C. M. Bhargavi, “Automated Answer Sheet Evaluation Using OCR and NLP,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 14, no. 1, pp. 45–52, 2025.

[2] J. Memon, M. Sami, R. A. Khan, and M. Uddin, “Handwritten Optical Character Recognition: A Comprehensive SLR,” IEEE Access, vol. 8, pp. 85465–85484, 2020.

[3] A. V. Gayer, Y. S. Chernyshova, and V. V. Arlazarov, “Recognition of Machine-Readable Zone in Identity Documents: A Review,” Pattern Recognition and Image Analysis, vol. 35, no. 2, pp. 245–267, 2025.

[4] J. Summaira, X. Li, et al., “Recent Advances and Trends in Multimodal Deep Learning: A Review,” Neural Computing and Applications, vol. 33, no. 15, pp. 8957–8985, 2021.

[5] F. F. Lubis, M. Mutaqin, et al., “Automated Short-Answer Grading Using Semantic Similarity Based on Word Embedding,” International Journal of Technology, vol. 12, no. 5, pp. 944–953, 2021.

[6] A. Ayaan and K.-W. Ng, “Automated Grading Using Natural Language Processing and Semantic Analysis,” MethodsX, vol. 10, pp. 102439, 2025.

[7] C. Garrido-Muñoz, A. Ríos-Vila, and J. Calvo-Zaragoza, “Handwritten Text Recognition: A Survey,” Pattern Recognition Letters, vol. 180, pp. 75–90, 2025.

[8] M. A. Rahman, M. T. Hasan, et al., “uFOIL: An Unsupervised Fusion of Image Processing and Language Understanding,” Pattern Analysis and Applications, vol. 28, pp. 1387–1401, 2025.

[9] A. Kiran, A. M. Poojari, and V. E. Salis, “Review on Smart Evaluation of Descriptive Answer Sheets,” International Journal of Innovative Research in Computer and Communication Engineering, vol. 12, no. 2, pp. 232–240, 2024.

[10] W. Xu, R. Mahmud, and W. L. Hoo, “Are Automated Essay Scoring Systems Competent in Real-Life Education Scenarios? A Systematic Literature Review,” Education and Information Technologies, vol. 29, pp. 1291–1313, 2024.

[11] N. Khan, A. A.-H. Ab Rahman, S. M. Rizvi, and S. A. Khan, “Systematic Literature Review of Machine Learning Models and Applications for Text Recognition,” Journal of Intelligent Systems, vol. 34, no. 1, pp. 55–74, 2025.

[12] M. A. Hearst, “The Debate on Automated Essay Grading,” IEEE Intelligent Systems, vol. 15, no. 5, pp. 22–37, 2020.

[13] M. Kaya and I. Cicekli, “A Hybrid Approach for Automated Short Answer Grading,” Knowledge-Based Systems, vol. 299, pp. 111042, 2024.

[14] A. Kiran, A. M. Poojari, and V. E. Salis, “Review on Smart Evaluation of Descriptive Answer Sheets,” International Journal of Innovative Research in Computer and Communication Engineering, vol. 12, no. 2, pp. 232–240, 2024.