International Advance Journal of Engineering Research (IAJER)
*Volume 09, Issue 07 (**July - 2026**), PP 01-10*
ISSN: 2360-819X
www.iajer.com

Research Paper **Open Access**

Automating the Evaluation of Non-Linear Student Trajectories: A Deep Learning Framework for Interactive Assessments.

Rachel C. Bual[1]
[1]*Manufacturing Engineering Department, Bulacan State University, Malolos, Bulacan*

ABSTRACT: Outcome-based Education (OBE) emphasizes measurable learning outcomes aligned with practical skills, incorporating frameworks like the Revised Bloom's Taxonomy (RBT) to cultivate higher-order thinking skills (HOTS) such as critical analysis and creativity. However, manually evaluating complex student assessments such as hazard identification in image tasks is time-consuming, subjective, and prone to inconsistencies, particularly with large student submissions. Traditional automated tools, like multiple-choice graders, fail to assess HOTS, leaving educators overwhelmed by administrative burdens. This study proposes a deep learning (DL) approach to automate the scoring of interactive student assessments within an OBE-RBT framework. Focusing on a boxed visual hazard identification activity, the research leverages DL models trained on 86 student submissions from a manufacturing engineering Occupational Safety and Health (OSH) course. Images were preprocessed, annotated for correct hazard identifications, and split into training (60%), validation (20%), and testing (20%) sets using Roboflow. The model achieved promising results, with precision and recall 97.7% and 98.8%, respectively, and a mean Average Precision (mAP50) of 99% after 200 epochs, demonstrating robust boxed hazard detection. By automating scoring, this approach reduces faculty workload, enhances grading accuracy, and aligns with OBE's pedagogical goals. Unlike prior AI research focused on text or structured evaluations, this study innovatively targets HOTS within visual tasks, bridging a gap in educational assessment technology. The study's findings imply that deep learning has the aptitude to fundamentally change assessment practices in education, thereby freeing up educators to allot more time to student mentorship, research, and extension activities, while simultaneously enabling the integration of artificial intelligence into pedagogical approaches.

Keywords: EduTech, object detection, learning assessment, Roboflow, OBE, RBT, deep learning

I. INTRODUCTION

Student-focused educational techniques have undergone numerous changes throughout the years as a result of the implementation of OBE. Previously, teaching was seen primarily as a means of imparting information to students. However, with OBE, emphasis is placed on ensuring that all students have reached clearly articulated and measurable learning outcomes [1, 2]. The OBE educational model is grounded on an assumption that every student should have developed a set of transferable skills that can be applied to the issues they encounter in the world. Some of these skills include critical thinking, problem-solving, and creativity [3]. It was through the use of the Revised Bloom's Taxonomy (RBT) that OBE was able to give educators a clearer definition by providing them with an explicit framework for organizing the cognitive domain of learning in order from basics (e.g., remembering; understanding) to advanced (e.g., analyzing; evaluating; creating). Once the outcome of a particular OBE course can be directly related to an RBT cognitive domain, the educator is then best able to create assessments that will encourage students to develop higher-order thinking skills (HOTS) [4] and transition from simple recall of information to innovative and valuable application of the same.

Yet the integration of OBE and RBT frameworks brings added layers of complexity for assessment processes. Faculty members' responsibility encompasses the assessment of multifaceted student assignments such as research papers, programming, and design portfolios using an extensive rubric pairing a designated taxonomy with a set of thorough evaluation criteria. For example, this research assesses the scoring of visual identification activities or quizzes based upon students' ability to identify hazards depicted in a given scenario or image.

The Revised Bloom's Taxonomy (RBT) is a systematic framework for helping develop an assessment of learning outcomes [5]. Despite this, the manual assessment of RBT competencies is labor intensive, time-

consuming (taking hours/days rather than minutes), sometimes subjective and can be prone to inconsistency, particularly when assessing a large number of students at once. Similarly, faculty have tremendous responsibilities, including instruction, developing instructional materials, performing research, engaging in extension activities and consulting with students. This can lead to faculty exhaustion as they struggle to manage their large number of responsibilities which may result in inaccurate evaluations or errors due to workload [6, 7]. Although traditional automated tools such as multiple choice question graders have been used to alleviate this burden, they do not provide any solutions to issues related to the inability to score open-ended tasks or evaluate higher-order cognitive skills. As a result, faculty members continue to have a heavy burden of administrative duties, which reduces the amount of time available to devote to individualised instruction and curriculum design.

The purpose of this research is to use deep learning (DL) to connect the goals of an outcomes-based education (OBE) system with the reality of how assessments are currently performed. The study uses rubrics aligned with the revised Bloom's Taxonomy (RBT) to evaluate students' work through trained DL models, which can automate the identification of hazards within the student responses. The importance of this research is that it can identify hazards (as accurate) from the actual responses of students, combining the power of state-of-the-art technology with a base of educational theory. Automating the scoring system provides equitable grading and allows faculty more time to devote to their other responsibilities, such as mentoring, research, and extension. Additionally, by incorporating DL architecture into the evaluation of student work, the study is further developing an A.I. system that will integrate with educational assessment methods.

II. REVIEW OF RELATED STUDIES

This section discusses relevant studies pertaining to the goal of this research. It is divided into discussions of oral assessments, essay and writing assessments, bubble sheet and multiple-choice question (MCQ) assessments, Automatic Short Answer Grading (ASAG), translation scoring, and figurative creativity test assessments.

New technologies for assessing Educational English as a Second Language (ESL) learners have begun to emerge throughout the world using artificial intelligence (AI). The research of Ji (2024) has introduced an automated method of assessing oral English using the BERT and Bi-LSTM (Bidirectional Long Short Term Memory) architectures via an integrated system called the BERTOphone, which looks at both linguistic and contextual aspects of language [8]. The work of Amalia et al. (2025) also developed an automated means of assessing oral responses in the Indonesian language using the BERT architecture and demonstrated that their system also has the capacity to assess other languages for their semantic complexity [9]. The research of Del Gobbo et al. (2023) created an improved means of computer-assisted essay grading (CAE) using both semantic-based (BERT) and lexical-based (TF-IDF) features through a hybrid approach to feature engineering [10]. In addition, Sree Lakshmi et al. (2024) evaluated short answer responses automatically using machine learning through a method they called Automatic Short Answer Grading (ASAG) that was scalable in relation to the number of short answer responses submitted [11]. Finally, the research of Süzen et al. (2020) introduced a two-stage process for evaluating written short answer responses. An unsupervised k-means clustering method was employed to categorize students' responses into proficiency levels based on the student responses being preprocessed (i.e., stemming, normalization), while supervised regression was utilised to optimise the prediction of grades using Hamming distance and MSE minimisation and providing proficiency-based feedback for each cluster [12].

In terms of their research efforts, essays, Filho et al., have made significant contributions to essay-scoring through developing automated essay scoring [13], while Da Corte and Baptista used natural language processing and machine learning to design automatic writing evaluation systems focusing on linguistic analysis [14]. In another new direction, to assess essays, Pasaribu et al. created GoogLeNet convolution neural network to evaluate essays using visual and textual features instead of typical natural language processing techniques [15]. For constructed responses in science, Latif and Zhai adjusted GPT-3.5-Turbo to preprocess and augment data in a way that will enable integration into JSON to improve the accuracy of scoring constructed responses [16]. Also focusing on improving analytical depth, Bewersdorff, Dunn, Feigel, McMillan, and O'Shaughnessey use natural language processing capabilities of both GPT-3.5 and GPT-4, as well as chain of thought prompting and prompt role-based techniques, to identify errors made by students while conducting experiments [17]. Finally, given the scarcity of data, Wang, Qu, Li, and Wong used GPT-3.5 to synthesize student essays to create additional training data for a multilingual bi-directional Encoder Representations from Transformers model, which demonstrated superior performance compared to self-augmentation methods in various categorization tasks [18].

In recent years, there have been new artificial intelligence (AI) advancements being made in the area of education assessment where the grading process can be automated through the application of computer vision and machine learning techniques [19]. Examples include Canny Edge Detection, Contour Detection and Image Segmentation techniques which were evaluated by Wata and Villaverde (2023) for processing exam sheet files in a manner that allows precise identifying of answer regions [19]. Rababaah (2025) and Ascencio et al. (2021) also discussed the use of Digital Image Processing (DIP) [20] and OpenCV [21], respectively, as ways to automatically grade multiple-choice questions (MCQs) thereby achieving more accurate detection of incorrectly marked items. Challenges associated with high volume exam grading, including human error and security issues (e.g., student impersonation), were also addressed through the use of Optical Mark Recognition (OMR) technology by Calado et al. (2019) at a university with approximately 30,000 exams per year, and by developing a Python-based OMR application using OpenCV [22]. In addition, Espinosa and Rangel (2022) used a combination of traditional machine learning (SVM and K-Wise) and deep learning techniques (CNN) to evaluate exams with an accuracy of 98% resulting from development and utilization of preprocessing techniques (binarization and OCR) and improved training data, thereby minimizing subjectivity of manual grading systems [23]. Finally, the hybrid approaches used by Espinosa and Rangel (2022) provided for the development of computer vision modules to be utilized in both the CFG method of segmentation and the MCS method of identifier detection thereby improving overall accuracy when used in conjunction with other machine learning algorithms [23].

Moreover, AI is transforming language assessment by automating subjective grading tasks. Wu et al. (2022) developed an efficient translation scoring system using BERT for accuracy (via cosine similarity) and GPT-2 for fluency (through probability scoring), demonstrating how NLP can ensure objective, scalable evaluations while reducing manual workload [24].

Furthermore, AI's recent impact on educational assessment is evident in the innovative methods used to evaluate various skills. Acar et al. (2025) developed the Ocsai-D system for figural creativity scoring [25], finding that supervised learning with vision transformers (particularly BEiT) outperformed unsupervised methods due to superior scalability and transfer learning capabilities. Their work briefly highlights AI's increasing potential to assess complex cognitive abilities like creativity, offering promising applications for standardized, large-scale educational evaluation.

Finally, the majority of researchers have focused on grading of automatically generated essays (or short answer) and translations and on scoring objective tests (MCQs or image-based assessments), leaving limited focus on grading or assessing higher level cognitive skills (higher order) and/or creative outputs. Although data and research demonstrate how well AI performs for linguistic and structured assessments, the data from this study provides a unique and innovative approach to assessing OBE-based activities located in the "understanding" level of the revised version of Bloom's Taxonomy, thereby filling an important gap in AI used for educational assessments. This study highlights the cognitive processes involved in assessing learning students' results as opposed to only evaluating their output and expanding AI's capability of measuring deeper levels of learning consistent with and supporting 21st century pedagogical practices.

III. MATERIALS AND METHODS

This part of the study discusses the materials used such as the activity in Occupational Safety and Health (OSH) course, and the framework used in the study to generate an object detection model.

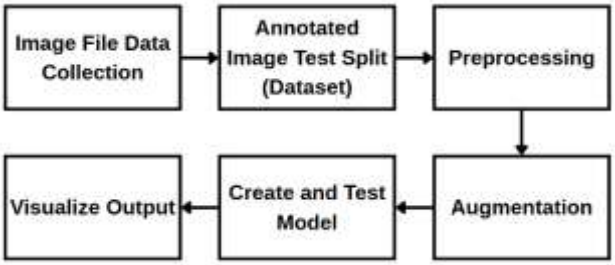

Fig 1. Methodological Framework utilized in the study

The students who submitted raw digital images are current and/or former manufacturing engineering students enrolled in or completed the OSH Course. Students received a digital version of their activities via PowerPoint, along with instructions detailing how the students could answer and a link to upload their image. Google Drive was the other delivery method. The research team compiled the responses provided to them from

these 86 responses. For reference, the image below (Figure 2) is also an example of what students received to create an image for submission. The AI transformed the image to a similar black and white or grayscale format prior to conducting the preprocessing element (again, to provide for consistency with the in-person assessment - blue box as default). The students were also given a pattern (box pattern) to use when submitting their images in response to their submitted activity. This will provide a streamlined process for the researchers to successfully and quickly focus on collecting these images by eliminating confusion about the relevant elements of the image.

Fig 2. Occupational Safety and Health (OSH) Activity [26]

In addition to the above, participants were instructed to capture an image of each of the maximum 14 hazard types that were observed with the goal of achieving tight bounds around the hazard, so that there was minimal area (too small or too large) surrounding each hazard to create maximum accuracy. Also, an example bounding box was included in the PowerPoint to assist with consistency in capturing images of hazards.

A deep learning methodology, Roboflow, was utilized as a mechanism for scoring the student's assessment/hazard identification activities. There were 86 raw images used as the input dataset and they were annotated and divided into 60:20:20 split for training, validation and testing respectively. In order to ensure that students were awarded points for their correct identification of hazards, this scoring scheme was based upon the number of hazards that the student correctly identified within the designated bounding box. During the annotation of each input image, only the correctly bound boxes were labelled using the annotation class name of 'correct' which ensured that all scoring was completed based upon the correct identification of hazards. The study did not annotate any bounding boxes that had incorrectly bounded boxes, in order to minimise the chance of confusion with the model.

The images were preprocessed within Roboflow through orientation and resizing to 1080 x 640 pixels. Images were then augmented through the following transformations: Hue (-50° to +50°), Saturation (-50° to +50°), Brightness (0 to +50%), and Exposure (-50° to +50%). Following the completion of the preprocessing and augmentation processes, the overall dataset consisted of 156 training images, 17 validation images, and 17 test images, for a total of 190 images.

IV. RESULTS AND DISCUSSION

A. Dataset Annotations

This subsection demonstrates the annotation method applied to the raw image files. Using the preferred blue box shown in Fig 2, the students replicated it to box each hazard they noticed.

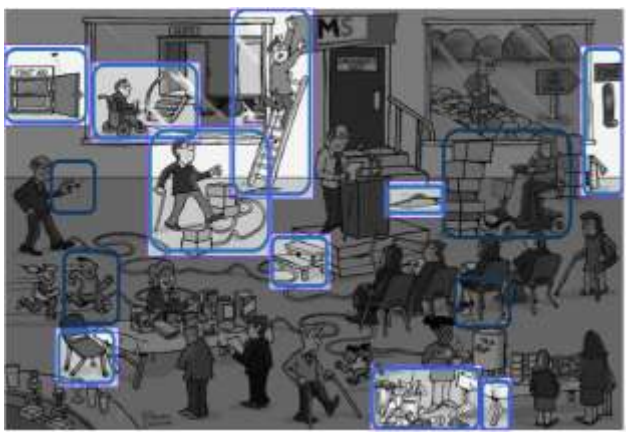

Figure 3. Annotated Images vs Boxed Hazards by Students

Moreover, from the assessor's or faculty's perspective, aligning with the goal of this study, the aim is to score the students' answers by detecting the hazards they have boxed. The study annotated only the correctly boxed hazards, while incorrect ones were not annotated. The accuracy performance is visualized in Figure 7, within the visualization subsection.

B. Train Losses

These plots show the loss values during training, which indicate how well the model is learning to detect boxed hazards (bounding boxes) on the training dataset.

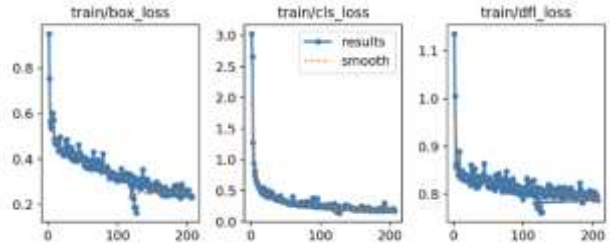

Fig 4. Model Graph: Train Losses

Box/Train_loss illustrates how well the model performs at predicting the bounding boxes of the object; it can be seen that the loss starts at ~0.8 and decreases rapidly to below ~0.2 after about 120 epochs; it then stabilizes at approximately 0.1. This indicates that the model has learned how to rapidly and accurately predict the location of the bounding boxes and to continue improving the predictions. Box/Train lines up with how well the model performs at classification by looking at train/cls_loss. It starts at approximately 3.0 and steadily decreases to around 0.5 by the 200th epoch. The steady decline leads us to believe that although the model is progressively getting better at classifying the hazards in the boxes around the students, the initial higher loss and slower convergence for the classification loss compared to the box loss may be indicative of how difficult it is to classify the hazards in the boxes due to the how variable the process and style is to label the objects in the dataset. The last measure of performance is represented by train/dfl_loss; this value starts at ~1.1 and decreases to about ~0.8 after 200 epochs. The pace of decline in this measurement is a bit slower than the box loss, indicating that the fine-tuning of the distribution of the bounding box predictions took longer to achieve, but the model did show continued improvement over time.

C. Validation Losses

These plots show the model's performance on the validation set of 17 images, which helps assess generalization to unseen data.

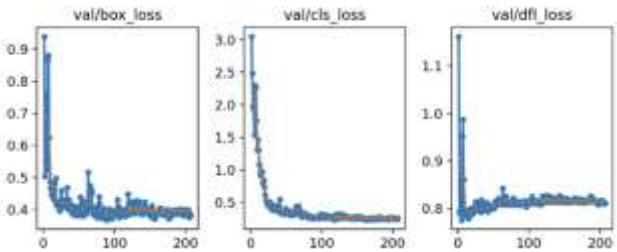

Fig 5. Model Graph: Validation Losses

The validation box loss is measured as the val/box_gain or val/cls_gain. The initial value of the validation box loss is 0.9 and it decreases to approximately 0.4 over the course of 200 epochs. The pattern of the validation box gain mirrors that of the box loss pattern, however, the rate of variability in the validation box gain is higher than in the box loss. The final value of 0.4 for the box gain indicates that the model generalizes to some extent to bounding box prediction.

Based on the larger difference in loss values between val/box and box, the potential for overfitting exists, especially when compared to the val/cls_loss, which starts at 3.0 and declines to about 0.5 after 200 epochs, again mirroring the training loss patterns for classification losses. Based on these patterns, the model generalizes well in classification tasks with minimal overfitting and the loss patterns of the training and validation

classification tasks for bounding objects are highly correlated, indicating a robust amount of learning for bounding objects within the context of hazard identification during the activity performed. Validation DL loss mirrors the DFL loss of the training data in terms of the initial value and decreasing from 1.1 to approximately 0.8. Clearly, validation DFL loss follows a trend similar to DL training indicating that the model generalizes well to the uncertainty associated with bounding objects, although the relatively small size of the validation dataset may be a contributing factor to the variability seen in validation DFL loss.

D. Metrics

These plots show evaluation metrics, likely for the validation set, to measure the model's performance in terms of precision, recall, and mean Average Precision (mAP).

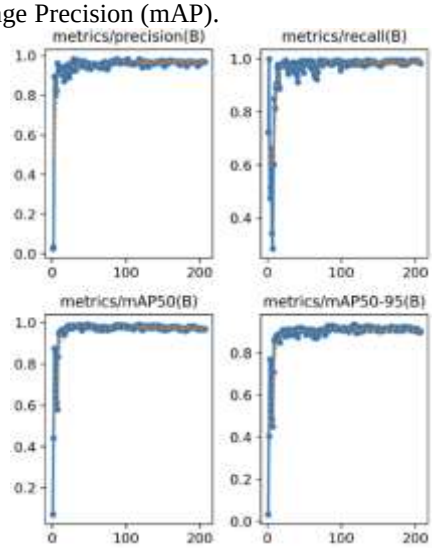

Fig 6. Model Graph: Metrics

The metrics/precision measuring the proportion of predicted bounding boxes that are correct (true positives / (true positives + false positives)) starts around 0.20 and goes up to about 0.90 by 200 epochs which indicates improved performance at identifying boxed hazards correctly and reduced false positives. The metrics/recall measuring the proportion of all boxed hazards that are correctly identified by the model (true positives / (true positives + false negatives)), has a starting point of 0.40 and also increases to about 0.80 by 200 epochs. This indicates that most of the boxed hazards in the images could be detected by the model; however, since the recall does not reach 1.00, some of these boxed hazards may have been missed. The metrics/mAP50, which is the Mean Average Precision at an IoU threshold of 0.50, evaluates the overall performance of the model in terms of its precision and recall—starting at 0.20 and increasing to approximately 0.90. This indicates that the model had high precision in identifying boxed hazards with an overlap, at a minimum, of 50% with ground truth. The metrics/mAP50-95 is a stricter measurement as the mAP score averaged over multiple IoU thresholds ranging from 0.50 to 0.95; with initial averages of approximately 0.10 and ending near 0.85, indicating decent but not optimal performance.

E. Visualization

The generated model was tested and visualized using the same raw image files gathered from the students. A sample of the object detection output is shown in Fig 6.

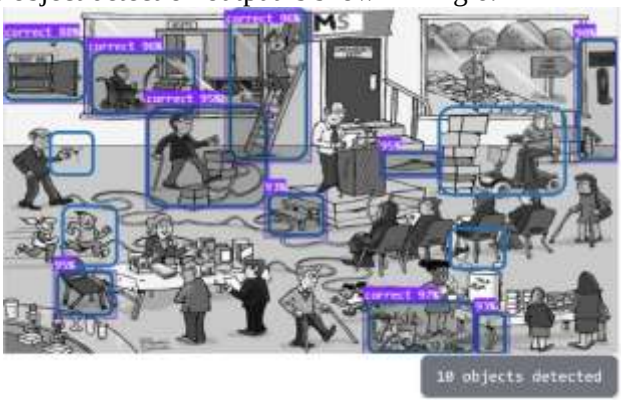

Fig 7. Sample visualized image in Roboflow using actual student's submission

The constructed model was able to identify ten objects, which were hazardous as per the boxes created by the student during the assessment. The model was also able to identify object areas that were incorrectly marked, as shown by the areas the model did not identify as objects. This illustrates how using deep learning for object detection can provide a quick and effective way to calculate a student's score on this type of learning assessment.

IV. CONCLUSION

In employing a constructivist approach to evaluate higher-order cognitive abilities of university students using an OBE model, this research provides evidence of how deep learning analysis (DL) can be used for online assessment and grading of interactive assessment tasks within the higher education environment. The DL model developed in this study for the digital display of 'boxer' hazards was able to detect with high accuracy (97.7%) and sensitivity (98.9%) and provided an overall mean average precision (mAP50 = 99%), thus providing a viable alternative to manual grading of assessment tasks. This innovative means of conducting assessments not only relieves the burden placed on faculty but also eliminates subjective grading and provides for fair assessment by solving many of the issues currently being confronted in the 21st-century education system.

Typical tools used in higher education only use objective assessment measures; however, the application of the DL approach to assessing higher-order cognitive abilities represents a transformation of the function of AI and its potential application in assessment in education. The impact of this research may extend beyond the identification of hazards within the 'boxer' environment, providing scalable solutions for complex and diverse assessments and supporting the OBE model by providing a mechanism to assess student progress and achievement of outcomes.

REFERENCES

[1]. D. Hariyani, P. Hariyani, S. Mishra, and M. K. Sharma, “A literature review on lean tools for enhancing the quality in the outcome-based education system,” *Thinking Skills and Creativity*, vol. 57, p. 101793, Sep. 2025, doi: 10.1016/j.tsc.2025.101793.

[2]. S. S. Shanto, Z. Ahmed, and A. I. Jony, “A proposed framework for achieving higher levels of outcome-based learning using generative AI in education,” *Educ. Technol. Q.*, vol. 2025, no. 1, pp. 1–15, Mar. 2025, doi: 10.55056/etq.788.

[3]. A. M. Almuhaideb and S. Saeed, “Fostering Sustainable Quality Assurance Practices in Outcome-Based Education: Lessons Learned from ABET Accreditation Process of Computing Programs,” *Sustainability*, vol. 12, no. 20, p. 8380, Oct. 2020, doi: 10.3390/su12208380.

[4]. K. Ravichandran and A. Virgin B, “Bloom’s Taxonomy Categories in the Economy of Literature Teaching-Learning Process,” *IRJMS*, vol. 05, no. 03, pp. 721–727, 2024, doi: 10.47857/irjms.2024.v05i03.0827.

[5]. E. Benli Özdemir, E. Yilmaz, and M. Selvi, “Comparative Analysis of the 2018 and 2024 Science Curricula about Environmental Topics Based on Bloom’s Taxonomy,” *Kastamonu Eğitim Dergisi*, vol. 33, no. 1, pp. 35–47, Jan. 2025, doi: 10.24106/kefdergi.1628228.

[6]. H. Ujir, S. F. Salleh, A. S. W. Marzuki, H. F. Hashim, and A. A. Alias, “Teaching workload in 21st century higher education learning setting,” *IJERE*, vol. 9, no. 1, p. 221, Mar. 2020, doi: 10.11591/ijere.v9i1.20419.

[7]. K. O’Meara, C. J. Lennartz, A. Kuvaeva, A. Jaeger, and J. Misra, “Department Conditions and Practices Associated with Faculty Workload Satisfaction and Perceptions of Equity,” *The Journal of Higher Education*, vol. 90, no. 5, pp. 744–772, Sep. 2019, doi: 10.1080/00221546.2019.1584025.

[8]. Y. Ji, “A Deep Learning Assessment Model for Oral Learning in English Online Education,” *Journal of Network Intelligence*, vol. 9, no. 2, pp. 896–917, May 2024.

[9]. A. Amalia, M. S. Lydia, M. A. Muchtar, F. Y. Manik, Sinu, and D. Gunawan, “Mitigating Bias and Assessment Inconsistencies with BERT-Based Automated Short Answer Grading for the Indonesian Language,” *IAENG International Journal of Computer Science*, vol. 52, no. 3, pp. 533–545, 2025.

[10]. E. del Gobbo, A. Guarino, B. Cafarelli, and L. Grilli, "GradeAid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation," *Knowl Inf Syst*, vol. 65, no. 10, pp. 4295–4334, Oct. 2023, doi: 10.1007/s10115-023-01892-9.

[11]. P. Sree Lakshmi, J. B. Simha, and R. Ranjan, "Empowering Educators: Automated Short Answer Grading with Inconsistency Check and Feedback Integration using Machine Learning," *SN COMPUT. SCI.*, vol. 5, no. 5, p. 653, Jun. 2024, doi: 10.1007/s42979-024-02954-7.

[12]. N. Süzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, "Automatic short answer grading and feedback using text mining methods," *Procedia Computer Science*, vol. 169, pp. 726–743, 2020, doi: 10.1016/j.procs.2020.02.171.

[13]. A. Filho, F. Concatto, H. Antonio Do Prado, and E. Ferneda, "Comparing Feature Engineering and Deep Learning Methods for Automated Essay Scoring of Brazilian National High School Examination:," in *Proceedings of the 23rd International Conference on Enterprise Information Systems*, Online Streaming, --- Select a Country ---: SCITEPRESS - Science and Technology Publications, 2021, pp. 575–583. doi: 10.5220/0010377505750583.

[14]. M. Da Corte and J. Baptista, "Leveraging NLP and Machine Learning for English (L1) Writing Assessment in Developmental Education:," in *Proceedings of the 16th International Conference on Computer Supported Education*, Angers, France: SCITEPRESS - Science and Technology Publications, 2024, pp. 128–140. doi: 10.5220/0012740500003693.

[15]. N. G. Pasaribu, G. Budiman, and I. D. Irawati, "Image-Based Essay Scoring Deep Learning Using a CNN Model GoogLeNet," in *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Yogyakarta, Indonesia: IEEE, Aug. 2024, pp. 328–333. doi: 10.1109/ICITISEE63424.2024.10730128.

[16]. . Latif and X. Zhai, "Fine-tuning ChatGPT for automatic scoring," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100210, Jun. 2024, doi: 10.1016/j.caeai.2024.100210.

[17]. Bewersdorff, K. Seßler, A. Baur, E. Kasneci, and C. Nerdel, "Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters," *Computers and Education: Artificial Intelligence*, vol. 5, p. 100177, 2023, doi: 10.1016/j.caeai.2023.100177.

[18]. K. Cochran, C. Cohn, J. F. Rouet, and P. Hastings, "Improving Automated Evaluation of Student Text Responses Using GPT-3.5 for Text Data Augmentation," in *Artificial Intelligence in Education*, vol. 13916, N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, and V. Dimitrova, Eds., in Lecture Notes in Computer Science, vol. 13916. , Cham: Springer Nature Switzerland, 2023, pp. 217–228. doi: 10.1007/978-3-031-36272-9_18.

[19]. M. G. Wata and J. F. Villaverde, "Bubble-Sheet Assessment Checker with Test Grader Using Computer Vision Through Raspberry Pi," in *2023 IEEE 6th International Conference on Computer and Communication Engineering Technology (CCET)*, Beijing, China: IEEE, Aug. 2023, pp. 137–142. doi: 10.1109/CCET59170.2023.10335127.

[20]. A. R. Rababaah, "Machine vision algorithm for MCQ automatic grading – MVAAG," *IJCVR*, vol. 15, no. 2, pp. 233–251, 2025, doi: 10.1504/IJCVR.2025.144786.

[21]. H. E. Ascencio, C. F. Pena, K. R. Vasquez, M. Cardona, and S. Gutierrez, "Automatic Multiple Choice Test Grader using Computer Vision," in *2021 IEEE Mexican Humanitarian Technology Conference (MHTC)*, Puebla, Mexico: IEEE, Apr. 2021, pp. 65–72. doi: 10.1109/MHTC52069.2021.9419920.

[22]. M. P. Calado, A. A. Ramos, and P. Jonas, "An application to Generate, Correct and Grade Multiple-choice Tests," in *2019 6th International Conference on Systems and Informatics (ICSAI)*, Shanghai, China: IEEE, Nov. 2019, pp. 1548–1552. doi: 10.1109/ICSAI48974.2019.9010132.

[23]. A. Espinosa and J. C. Rangel, "Digitalized Academic Exam Evaluation System, Using Deep Learning and Computer Vision," in *2022 8th International Engineering, Sciences and Technology Conference (IESTEC)*, Panama, Panama: IEEE, Oct. 2022, pp. 215–222. doi: 10.1109/IESTEC54539.2022.00040.

[24]. D. Wu, M. Wang, and X. Li, "Automatic Scoring for Translations Based on Language Models," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–10, Jun. 2022, doi: 10.1155/2022/2171206.

S. Acar, P. Organisciak, and D. Dumas, "Automated Scoring of Figural Tests of Creativity with Computer Vision," *Journal of Creative Behavior*, vol. 59, no. 1, p. e677, Feb. 2025, doi: 02/jocb.677.

[25]. [26] "Spot The Hazard | Mevzuisg," Pinterest. Accessed: Apr. 07, 2025. [Online]. Available: https://ph.pinterest.com/pin/951174383775115783/