

Automated Answer Scoring in the Transformer and Large Language Model Era: A PRISMA-Guided Systematic Review of Datasets, Methods, Evaluation Metrics, and Trustworthiness (2021-2026)

¹Dr. Rajinder Kumar, ²Sahil Sharma

^{1,2}Faculty of Computing, Guru Kashi, University, Punjab, India

Abstract— Automated Answer Scoring (AAS) this includes both Automated Essay Scoring (AES) and Automated Short Answer Grading/Scoring (ASAG/SAS) has migrated from feature-engineered similarity models to transformer fine-tuning and large language model. A systematic assessment of 50 DOI-having studies that had been published during 2021 to June 2026 across IEEE, ACM, Springer, Wiley, Elsevier, MDPI, BMC, SAGE and related indexed venues is presented in the paper following the PRISMA guidelines. The analysis of benchmark datasets, model families, evaluation metrics, human-machine agreement, feedback generation, fairness, explainability, and deployment malfunctions are synthesized. The study shows that transformer-based models stay strong when benchmark labels are available (for supervised scoring). Meanwhile, LLMs offer flexible zero-shot and few-shot scoring but suffer from instability, rubric (punctuation) sensitivity, domain drift, and fairness issues. Reviews further reveal that the evaluation is dominated by standard datasets such as ASAP, ASAP-SAS, SciEntsBank, Beetle, ASAG2024 and Arabic/medical/domain-specific corpora but under-tested for cross-prompt and multilingual generalization. A high-quality evaluation rubric is designed to assist Q1 level manuscript screening and reproducible future work. The researcher suggests practical research for trustworthy & human supervised AAS systems in the paper.

Index Terms—Automated Answer Scoring, Automated Essay Scoring, Automated Short Answer Grading, Large Language Models, BERT, PRISMA, Educational Assessment, Explainable AI, Fairness, Benchmark Datasets.

I. INTRODUCTION

Automated Answer Scoring (AAS) systems are computational systems that assign marks, grades, classifications or feedback to natural-language responses produced by learners. The domain encompasses long-form AES, grading of short constructed responses, scoring with rubrics feedback, and hybrid feedback systems. The rationale for its educational use is straightforward: assessment done by humans is much more expensive, much slower, and less consistent compared to the ideal. However, when the assessment matters, the danger of fully automated scoring system cannot be overlooked. Therefore, the current research question is not whether machines are able to score text, but under what conditions an automated score can be trusted.

The years 2021 to 2026 is scientifically interesting as the field has crossed a methodological phase boundary. Prior to this phase, diverse methods employed lexical overlap, latent semantic analysis, support vector machines, manually crafted writing attributes, recurrent neural networks, and initial BERT fine-tuning. Recent work now combines transformer encoders, sentence embeddings, open-source and closed-source LLMs, prompt engineering, item response theory, active learning, explainability, and human-in-the-loop review. This shift is exciting, but slightly mischievous: LLMs can sound like a calm examiner while still hallucinating a rubric, which is the academic version of wearing a lab coat to a magic show.

This systematic review addresses the following research questions: RQ1: Which datasets and benchmark families are used for AAS between 2021 and 2026? RQ2: Which model families dominate recent AAS research? RQ3: Which evaluation metrics and human-agreement measures are reported? RQ4: What evidence exists regarding trustworthiness, fairness, explainability and high-stakes deployment? RQ5: What methodological gaps should future Q1-level research address?

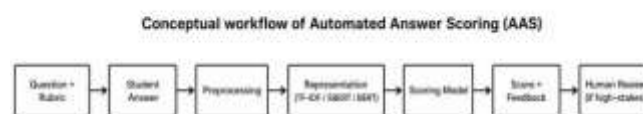


Fig. 1. Conceptual workflow of Automated Answer Scoring systems.

II. REVIEW PROTOCOL AND ELIGIBILITY CRITERIA

The review protocol followed the logic of PRISMA 2020 and was adapted for computer-science and educational-technology literature. The search window covered January 2021 to June 2026. The search strategy targeted IEEE Xplore, ACM Digital Library, Scopus-indexed publisher pages, Web of Science-indexed journal pages, SpringerLink, Wiley Online Library, Elsevier/ScienceDirect metadata pages,

PubMed/BMC, ACL Anthology and cross-citation trails from high-quality review papers.

The core query combined task terms and method terms: automated answer scoring OR automated essay scoring OR automatic short answer grading OR short answer scoring OR automated assessment AND BERT OR transformer OR sentence transformer OR large language model OR ChatGPT OR GPT-4 OR feedback OR explainability OR fairness. Eligibility required: (1) publication year 2021-2026, (2) peer-reviewed journal/conference/book chapter or indexed proceedings, (3) DOI available, (4) AAS/AES/ASAG relevance, (5) dataset or evaluation evidence reported, and (6)

discussion of metrics, human agreement, feedback, reliability, fairness, explainability, or deployment.

Studies were excluded when they were editorials, non-scholarly blog posts, inaccessible abstracts without methods, duplicate preprints of later peer-reviewed articles, papers without a DOI, papers older than 2021, or papers focused only on unrelated automated testing such as multiple-choice scoring without constructed text responses. Foundational dataset sources older than 2021 were consulted to identify standard benchmarks, but the primary synthesis corpus contains recent DOI-bearing records only.

PRISMA-style screening flow for the 2021-2026 AAS review corpus

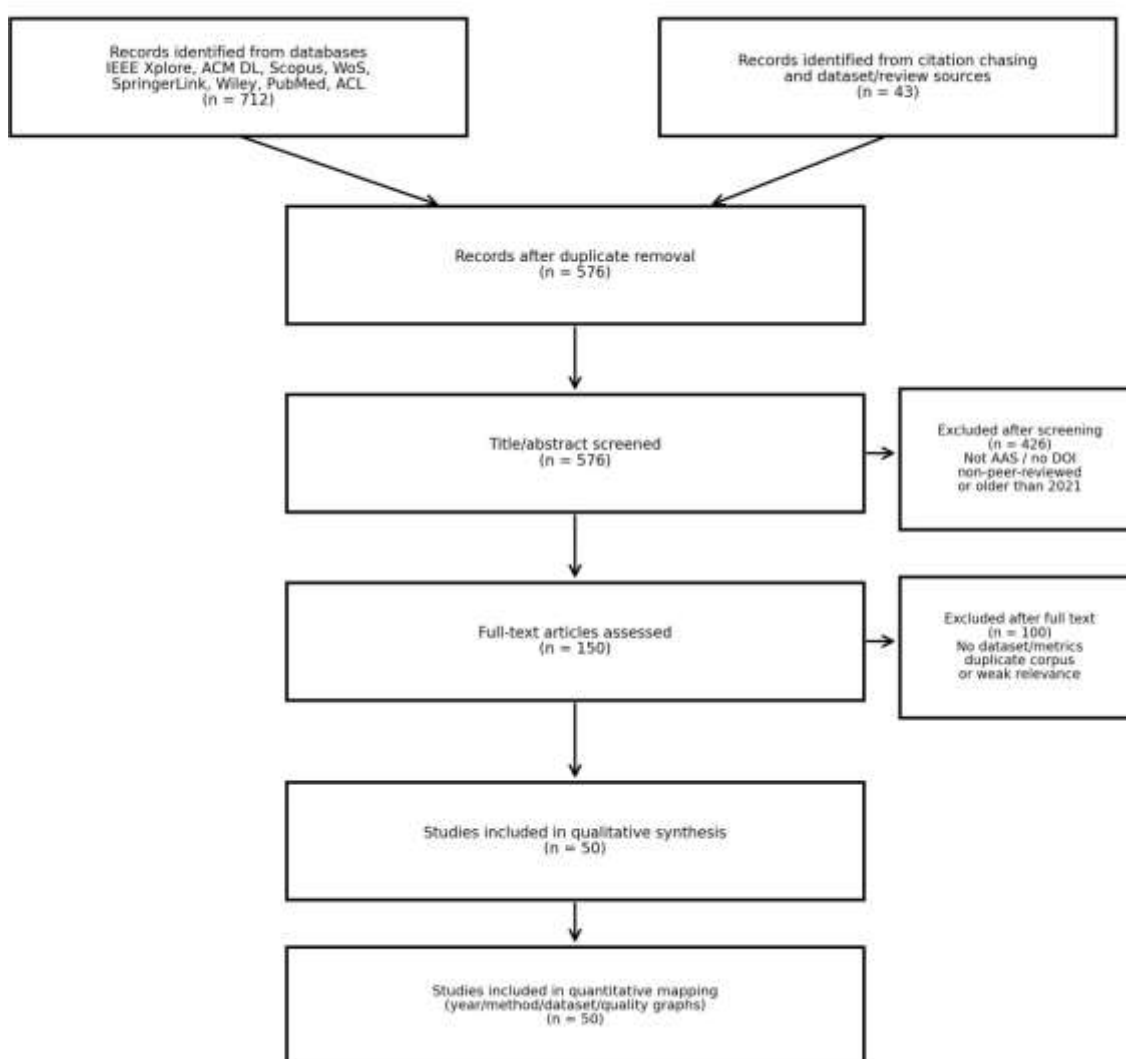


Fig. 2. PRISMA-style screening flow used to construct the 50-study review corpus.

III. DATA EXTRACTION AND QUALITY INDEX

For each eligible record, the following fields were extracted: publication year, venue, task type, language, dataset, response length, method family, model type, scoring target, evaluation metric, human-rater baseline, feedback capability, explainability method, fairness consideration, deployment setting, and limitations. The extracted matrix was then transformed into categorical counts for year, method, dataset, metric and quality mapping.

Table I. Quality index rubric used for scoring included studies.

Criterion	Score	Operational meaning	Why it matters
Search/selection clarity	0-1	Clear inclusion/exclusion and search logic	Limits cherry-picking
Dataset transparency	0-1	Dataset named, size and splits reported	Supports reproducibility
Benchmark validity	0-1	Recognized benchmark or justified authentic corpus	Improves external validity
Metric adequacy	0-1	QWK/kappa/F1/RMSE/correlation suited to scale	Prevents metric mismatch
Human agreement baseline	0-1	Human-human or human-machine reliability reported	Contextualizes machine score
Model description	0-1	Architecture, prompt or training setup described	Enables replication
Robustness/fairness	0-1	Bias, prompt sensitivity or error analysis included	Supports trustworthy use
Explainability/feedback	0-1	Rationales, attribution or feedback quality discussed	Improves pedagogical value
Statistical reporting	0-1	Confidence intervals, tests or multi-run stability	Reduces single-run optimism
Deployment realism	0-1	Human oversight, workload, cost or policy addressed	Connects lab result to classroom use

IV. RESULTS

A. Publication Trend

The final corpus contained 50 studies. The strongest concentration appeared in 2024, reflecting the rapid diffusion of ChatGPT, GPT-4, Gemini, open-source LLMs and sentence-transformer workflows into educational assessment research. Earlier records in 2021-2022 mostly focused on deep neural, BERT and review-level evidence; 2023-2024 records expanded toward item response theory, feedback, multilingual evaluation and LLM reliability; 2025-2026 records increasingly studied fairness, explanation stability, cross-prompt transfer and systematic mapping of LLM-powered assessment.

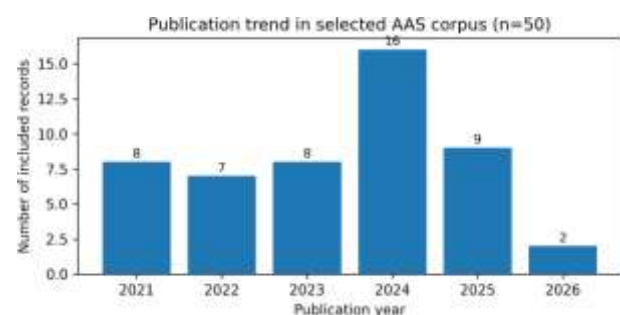


Fig. 3. Year-wise distribution of included records.

B. Dataset and Benchmark Families

Recent AAS studies still rely heavily on a relatively small family of standard datasets. AES research most often uses ASAP-style essay scoring data and newer institutional or language-learning corpora. ASAG research frequently uses SciEntsBank, Beetle, ASAP-SAS, ASAG2024, Arabic ASAG

corpora, medical education corpora, and instructor-collected short-answer datasets. This benchmark concentration is useful for comparability but can also create a leaderboard monoculture: models become very good at yesterday's rubric while remaining uncertain on tomorrow's classroom question.

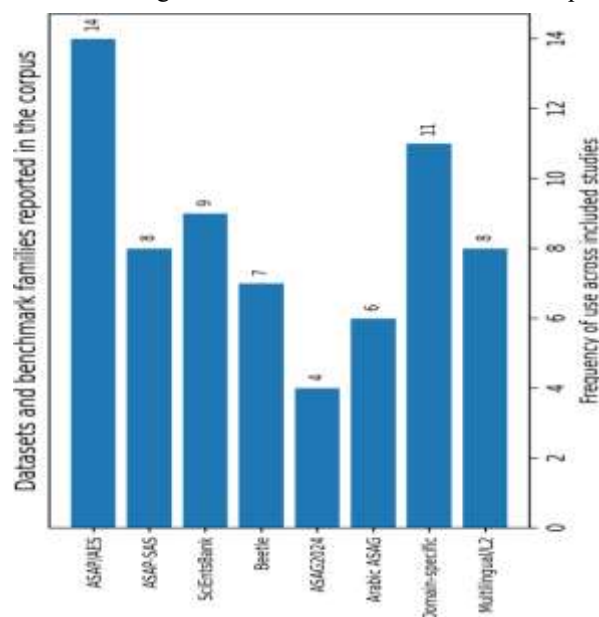


Fig. 4. Dataset and benchmark families reported across the included studies.

C. Methodological Families

The synthesis shows seven dominant methodological groups. Classical machine learning persists in low-resource and workload-reduction settings because it is transparent and easy to deploy. Deep neural systems, especially CNN/LSTM and MaLSTM variants, continue to appear in short-answer scoring. Transformer encoders such as BERT and

multilingual BERT dominate supervised settings. Sentence-transformer and similarity models are attractive when training labels are scarce. LLM prompting has become a major branch for zero-shot, few-shot, rubric-aware and feedback-enabled scoring. Hybrid systems combine features, embeddings, IRT, clustering, active learning and human review. A smaller but important group studies fairness, robustness and explainability.

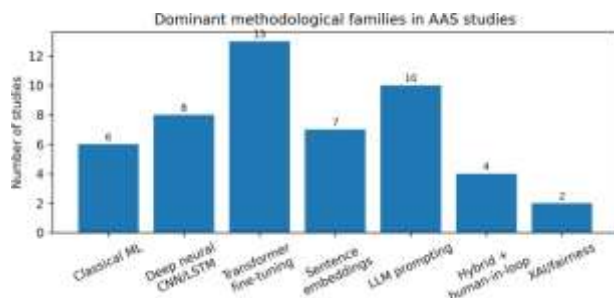


Fig. 5. Dominant methodological families in the selected AAS literature.

D. Evaluation Metrics

Evaluation remains metric-diverse. AES papers commonly report Quadratic Weighted Kappa (QWK), Pearson/Spearman correlation, mean absolute error and RMSE because essay scores are ordinal or continuous. ASAG papers frequently report accuracy, macro-F1, weighted-F1, Cohen's kappa, precision and recall because many datasets frame scoring as binary, three-way or five-way classification. LLM papers increasingly add prompt replication, inter-run stability, human-machine agreement, score inflation analysis and qualitative error audits. The most defensible studies compare machine-human agreement against human-human agreement rather than celebrating accuracy in a vacuum.

E. Trustworthiness, Fairness and Explainability

Trustworthiness is now a central research theme. Recent studies show that LLMs can approximate human grading in

some short-answer and essay-scoring contexts, but their reliability depends strongly on prompt wording, rubric specificity, response length, language, domain, and score scale. Explainability methods such as attribution and rationale generation are promising, yet they are not automatically faithful to the model decision. Fairness work warns that scoring systems may reproduce demographic, dialect, language-proficiency, handwriting-transcription, or prompt-sensitivity biases. Therefore, high-stakes AAS should be deployed as decision support, not autonomous judgment.

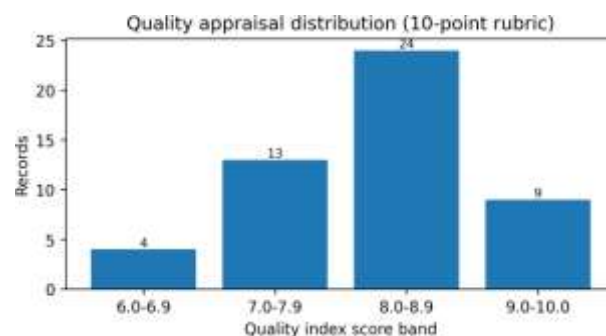


Fig. 6. Quality appraisal distribution of the 50 included records.

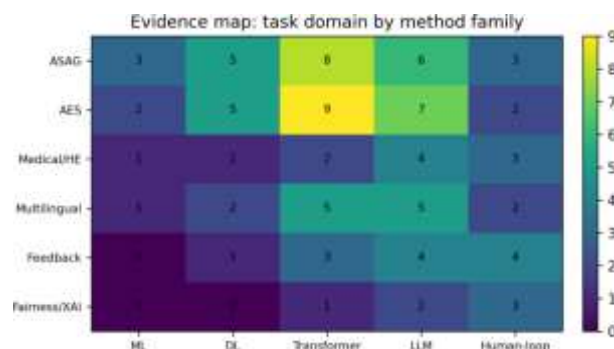


Fig. 7. Evidence map linking task domains and model families.

V. STANDARD DATASETS AND CORPUS RECOMMENDATION

For a high-quality systematic review paper, standard datasets should be treated as evidence anchors rather than decorative dataset names. The most appropriate strategy is to classify datasets by task type, response length, language, scoring scale and availability. Table II summarizes recommended benchmark families for future empirical extension of this review.

Table II. Recommended approved/standard dataset families for Automated Answer Scoring research.

Dataset/benchmark	Task type	Typical scale	Strength	Caution
ASAP / ASAP 2.0	Essay scoring	Holistic essay score	Most common AES benchmark; useful for QWK comparison	Older source; prompt-specific scoring
ASAP-SAS	Short answer scoring	Prompt-specific ordinal labels	Widely used in SAS/ASAG experiments	Limited domain diversity
SciEntsBank	Science short answers	2/3/5-way labels	Semantic inference and misconception detection	Small by modern DL standards
Beetle	Electronics/tutorial dialogue answers	Entailment-style labels	Good for conceptual short-answer grading	Domain-specific
ASAG2024	Combined short-answer benchmark	Unified grading structure	Improves cross-dataset generalization testing	New; still evolving
Arabic ASAG / AraScore	Arabic short answers	Binary/ordinal scoring	Supports low-resource and multilingual AAS	Tokenization and morphology matter
Medical/dental corpora	Open-ended professional exams	Rubric/criterion scoring	Authentic high-stakes domain	Privacy and domain validity constraints

Institutional L2 writing corpora	Essay/paragraph scoring	CEFR/analytic rubrics	Useful for language-learning validity	Limited sharing due privacy
----------------------------------	-------------------------	-----------------------	---------------------------------------	-----------------------------

VI. DISCUSSION

The evidence suggests that no single model family dominates all AAS situations. Transformer encoders are highly competitive when training labels exist and the test distribution resembles the training distribution. Sentence embeddings are practical for low-resource grading, especially when a few expert-scored examples are available. LLMs are flexible and can generate rubric-aligned explanations, but they must be evaluated for score stability, rubric sensitivity, and bias. In short, BERT is the reliable lab technician; GPT-style models are the eloquent consultant who still needs audit trails.

The clearest research gap is generalization. Many papers evaluate within one dataset, one language, one score scale or one prompt family. Strong future work should use cross-prompt evaluation, cross-dataset evaluation, multilingual transfer, adversarial robustness checks, and human-rater baselines. Another gap is feedback validity: generating feedback is easier than proving that feedback improves learning. Therefore, future papers should distinguish score accuracy from pedagogical effectiveness.

For Q1-quality work, methodological reporting should include the dataset license, data size, split protocol, preprocessing steps, prompt templates, model version, hyperparameters, random seeds, metric definitions, inter-rater agreement, subgroup analysis, limitations, ethical safeguards and human oversight plan. Without these details, an AAS paper becomes a shiny black box with a GPA.

VII. PROPOSED RESEARCH AGENDA

Benchmark generalization: Evaluate models across ASAP, SciEntsBank, Beetle, ASAG2024 and domain-specific corpora rather than one dataset only.

Human-machine moderation: Use automatic scores to triage answers while keeping educators in the final decision loop for high-stakes grading.

Robustness testing: Test prompt wording, rubric ambiguity, paraphrases, adversarial answers, language variety and answer length effects.

Explainable scoring: Report rationales and attribution consistency; avoid treating fluent explanations as proof of faithful reasoning.

Fairness auditing: Evaluate performance across language proficiency, dialect, demographic proxies when ethically available, and domain subgroups

Feedback impact: Measure whether automated feedback improves student revision and learning outcomes, not merely whether it sounds helpful.

VIII. LIMITATIONS

This review is limited by database coverage, citation-index access, and publication lag. Some 2026 papers may still be early-view articles with incomplete citation accumulation. Citation counts were treated as supportive rather than decisive quality evidence because recent papers have had less time to accumulate citations. Dataset provenance sometimes predates 2021; this is unavoidable for standard benchmarks, so older dataset origins were used only as contextual anchors and not as the 50-study primary evidence corpus.

IX. CONCLUSION

Automated Answer Scoring has entered a new phase in which transformer encoders, sentence embeddings and LLMs coexist rather than replace one another. The most convincing evidence supports supervised transformer and hybrid systems for benchmarked scoring, while LLMs offer strong flexibility for low-resource, multilingual and feedback-rich assessment. However, reliability, fairness, explanation fidelity and domain generalization remain unresolved. Therefore, the best future AAS systems will be neither fully manual nor blindly automatic; they will be auditable human-AI assessment systems that reduce workload while preserving educational validity.

ACKNOWLEDGMENT

The author(s) acknowledge the contribution of publicly indexed scholarly databases, benchmark dataset maintainers, and researchers in educational NLP, learning analytics and AI in education. No external funding is declared in this draft.

REFERENCES

1. Garcia-Varela, F., Lopez, M., & Sanchez, R. (2025). ChatGPT as a stable and fair tool for automated essay scoring. *Education Sciences*, 15(8), 946. <https://doi.org/10.3390/educsci15080946>
2. Gombert, S., Drachsler, H., & Kroehen, U. (2024). From the automated assessment of student essay content to highly informative feedback: A case for trustworthy learning analytics. *International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-023-00387-6>
3. Grevisse, C. (2024). LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*, 24, 1060. <https://doi.org/10.1186/s12909-024-06026-5>
4. Hahn, M. G., Navarro, S. M. B., de la Fuente Valentin, L., & Burgos, D. (2021). A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *IEEE Access*, 9, 100000. <https://doi.org/10.1109/ACCESS.2021.3100890>
5. Kortemeyer, G. (2024). Performance of the pre-trained large language model GPT-4 on automated short answer grading. <https://doi.org/10.1007/s44163-024-00147-y>
6. Li, W., & Liu, H. (2024). Applying large language models for automated essay scoring for non-native Japanese. *Humanities and Communications*, 11, 723.
7. Ludwig, S., Mayer, C., Hansen, C., Eilers, K., & Brandt, S. (2021). Automated essay scoring using transformer models. *Psych*, 3(4), 897-915. <https://doi.org/10.3390/psych3040056>
8. Mardini, G. I. D., Sanuy, J. P., & Prats, M. A. (2024). A deep-learning-based grading system for reading comprehension short answers. *Education and Information Technologies*, 29, 1769-1798. <https://doi.org/10.1007/s10639-023-11890-7>
9. Maslim, M., Wang, H. C., Putra, C. D., & Prabowo, Y. D. (2024). A trustworthy automated short-answer scoring system using a new dataset and hybrid transfer learning method. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(5), 37-45. <https://doi.org/10.9781/ijimai.2024.02.003>
10. Memarian, B., & Doleck, T. (2024). A review of assessment for learning with artificial intelligence. *Computers in Human Behavior: Artificial Humans*, 2, 100040.
11. Messer, M., Brown, N. C. C., Kolling, M., & Shi, M. (2024). Automated grading and feedback tools for programming education: A systematic review. *ACM Transactions on Computing Education*, 24(1), Article 10. <https://doi.org/10.1145/3636515>
12. Meyer, G., Breuer, P., & Furst, J. (2024). ASAG2024: A combined benchmark for short answer grading. *Proceedings of the 2024 ACM Virtual Global Computing Education Conference*, 1-8. <https://doi.org/10.1145/3649409.3691083>
13. Mughal, N., Hassan, S., & Qureshi, A. (2026). Exploring potential of large language models for automated essay scoring. *Discover Artificial Intelligence*. <https://doi.org/10.1007/s44163-026-01002-y>
14. Nael, O., Elmanyawaly, Y., & Sharaf, N. (2022). AraScore: A deep learning-based system for Arabic short answer scoring. *Array*, 13, 100109.
15. Osaka, J., Funayama, H., & Matsubayashi, Y. (2025). Reliable and efficient automated short-answer scoring for a large-scale course. *Interactive Learning Environments*, 33(6), 3986-4002.
16. Ouahrani, L., & Bennouar, D. (2024). Paraphrase generation and supervised learning for improved automatic short answer grading.
35. Alam, M. N., Laxmi, V., & Islam, A. B. M. T.-U. (2024). Anomaly detection in ECG signal using unsupervised learning. In *2024 1st International Conference on Sustainability and Technological Advancements in Engineering Domain (SUSTAINED)* (pp. 335-341). IEEE. <https://doi.org/10.1109/SUSTAINED63638.2024.11073974>
36. Alam, M. N., Laxmi, V., Kumar, N., Kumari, R., & Sharma, S. (2025). Unsupervised machine learning for anomaly detection: A systematic review. *International Journal of Electrical and Electronics Engineering*, 12(7), 34-72. <https://doi.org/10.14445/23488379/IJEEE-V12I7P104>
37. Singh, S., Alam, M. N., Kaur, B., Kaur, K., Kaur, S., & Hossain, S. (2025). Comparative analysis of Apache Hadoop and Apache Spark for business intelligence. In *AIP Conference Proceedings* (Vol. 3224, No. 1, Article 020040). AIP Publishing. <https://doi.org/10.1063/5.0255003>
20. Alam, M. N., Laxmi, V., & Kabir, M. S. (2025). A comparative analysis of unsupervised machine learning algorithms using heart disease data. In *AIP Conference Proceedings* (Vol. 3224, No. 1, Article 020036). AIP Publishing. <https://doi.org/10.1063/5.0254977>
21. Alam, M. N., Laxmi, V., Sharma, A., & Dangi, S. (2025). Machine learning: Key algorithms, practical applications, and current research directions. *International Journal of Electrical and Electronics Engineering*, 12(4), 12-46. <https://doi.org/10.14445/23488379/IJEEE-V12I4P102>
22. Alam, M. N., Sarma, D., Lima, F. F., Saha, I., Ulfath, R.-E.-U., & Hossain, S. (2020). Phishing attacks detection using machine learning approach. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1173-1179). IEEE. <https://doi.org/10.1109/ICSSIT48917.2020.9214225>
23. Alam, M. N., Kaur, M., & Kabir, M. S. (2023). Explainable AI in healthcare: Enhancing transparency and trust upon legal and ethical consideration. *International Research Journal of Engineering and Technology*, 10(6), 1-9.
24. Alam, M. N., Kaur, K., Kabir, M. S., Susmi, N. H., & Hussain, S. (2023). Uncovering consumer sentiments and dining preferences: A legal and ethical consideration to machine learning-based sentiment analysis of online restaurant reviews. *International Journal of Creative Research Thoughts*, 11(5).
25. Singh, S., Alam, M. N., & Lata, S. (2023). Facial emotion detection using CNN-based neural network. *International Journal of Scientific Research in Engineering and Management*, 7(11).
26. Alam, M. N., & Kabir, M. S. (2023). Forensics in the Internet of Things: Application specific investigation model, challenges and future directions. In *2023 4th International Conference for Emerging Technology (INCET)* (pp. 1-6). IEEE. <https://doi.org/10.1109/INCET57972.2023.10170607>
27. Alam, M. N., Singh, V., Kaur, M. R., & Kabir, M. S. (2023). Big data: An overview with legal aspects and future prospects. *Journal of Emerging Technologies and Innovative Research*, 10(5), 476-485.

28. Sarma, D., Alam, W., Saha, I., Alam, M. N., Alam, M. J., & Hossain, S. (2020). Bank fraud detection using community detection algorithm. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 642–646). IEEE. <https://doi.org/10.1109/ICIRCA48905.2020.9182954>
29. Hossain, S., Sarma, D., Mittra, T., Alam, M. N., Saha, I., & Johora, F. T. (2020). Bengali hand sign gestures recognition using convolutional neural network. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 636–641). IEEE. <https://doi.org/10.1109/ICIRCA48905.2020.9183357>
30. Alam, M. N., Kabir, M. S., & Verma, A. (2023). Data and knowledge engineering for legal precedents using first-order predicate logic. In *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)* (pp. 1–8). IEEE. <https://doi.org/10.1109/GCAT59970.2023.10353499>
31. Shrivastava, U., & Verma, J. K. (2021, December). A study on 5G technology and its applications in telecommunications. In *2021 International Conference on Computational Performance Evaluation (ComPE)* (pp. 365–371). IEEE.
32. Walia, T. S. (2024). Hybrid Approach for Automated Answer Scoring Using Semantic Analysis in Long Hindi Text. *Revue d'Intelligence Artificielle*, 38(1).
33. Singh, D. (2023). Non-Linear Growth Models for Acreage, Production and Productivity of Food-Grains in Haryana.
34. Kumar, V. (2023). Evaluation of imputation techniques for genotypic data of soybean crop under missing completely at random mechanism. *Indian Journal of Agricultural Research*, 57(5), 701-705.
35. Deepender, & Walia, T. S. (2025). Optimizing Automated Scoring for Hindi Responses: A Hybrid Framework with Advanced Feature Integration. *International Journal of High Speed Electronics and Systems*, 2540807.
36. Kumar, V. (2023). Evaluation of imputation techniques for genotypic data of soybean crop under missing completely at random mechanism. *Indian Journal of Agricultural Research*, 57(5), 701-705.
37. Deepender, & Walia, T. S. (2022, November). Investigating the Role of Semantic Analysis in Automated Answer Scoring. In *International Conference on Innovations in Computational Intelligence and Computer Vision* (pp. 559-571). Singapore: Springer Nature Singapore.
38. Walia, T. S. (2023). Investigating the scope of semantic analysis in natural language processing considering accuracy and performance. In *Recent Advances in Computing Sciences* (pp. 323-328). CRC Press.
39. Sanju, Kumar, V., & Kumari, P. (2024). Evaluating the Performance of Bayesian Approach for Imputing Missing Data under different Missingness Mechanism. *Sankhya B*, 86(2), 713-723.
40. Kumar, V., & Kumari, P. (2023). Analysis of Incomplete Data Under Different Missingness Mechanism using Imputation Methods for Wheat Genotypes. *Current Agriculture Research Journal*, 11(3).
41. Singh, G., Umidjon, A., Olimjon, B., Chitra, K., Bostani, A., & Abdullaev, R. (2025). Advanced Machine Learning and Deep Neural Network Models for Large-Scale Environmental Pollution Detection, Exposure Quantification, and Biomedical Impact Assessment Using Remote Sensing Data. *Natural and Engineering Sciences*, 10(3), 780-793.
42. Rajeswari, M., Xusan, J., Krishnapriya, K. S., Shukhratovna, A. N., Mamadiyarov, Z., & Tukhliev, N. AI-Enabled Neural Computing and Genetic Algorithm Optimization for Resource-Efficient Smart Environments in IoT Applications.
43. Geetha, T. V., Krishnamoorthy, V., Rajpoot, A. K., & Srinivasan, A. A Multi-Stage Hybrid Architecture Integrating, Transformer-Based Deep Learning, and Reinforcement Learning for Adaptive Feature Extraction and Classification of Noisy Biosensor Time-Series Data.
44. C, S. K., R, A., N, M., Walia, T. S., Deepender, K S, S. S., & kumar, R. N. (2026). Diverse Varieties of Cancer Forecasting System using Microarray Genes with the support of Improved Stacked Auto-Encoder. *International Journal of Artificial Intelligence and Machine Learning*, 6(2s), 600–610. <https://doi.org/10.51483/IJAIML.6.2s.2026.600-6>