Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 14
TPCEE 2026

A Comparative Analysis of CNN, LSTM, and Transformer Models in AI-Enabled Online Education Platforms

Zhengshu Wang [1, *]

[1] College of Engineering, Stony Brook University, 11790 Long Island, United States

* Corresponding Author Email: zhengshu.wang@stonybrook.edu

Abstract. Artificial intelligence is being applied more and more broadly in online education to offer personalized learning paths, automatic grading, summary functions, etc. This narrative review compares convolutional neural networks (CNNs), long short-term memory networks (LSTMs) and Transformer-based models on the above tasks. Task suitability and deployment conditions are considered over reported accuracy values, rather than being used as a general reference. Based on the above research, CNNs are suitable for image and other inputs with local patterns; LSTMs are suitable when the order of learner behavior or language is required; and Transformer-based models are generally more appropriate for long-context semantic matching, summarization, open-ended feedback, retrieval and tutoring. Reported performance values cannot be directly compared between studies because tasks, datasets, languages, label distributions, hardware and evaluation metrics are all different. Model selection should therefore consider the predictive performance in conjunction with latency, resource requirements, privacy, fairness, explainability, factual accuracy and teacher supervision. No single Architecture is suitable for all educational environments. Educational value refers to how well a model's capacity suits the data structure, instructional goals, risk level and operating environment of the platform.

Keywords: CNN; LSTM; Transformer Models; Artificial intelligence; online education.

1. Introduction

Online education platforms generate diverse data at all times during learning, such as behavior logs, test results, learner-created content and multi-media submissions. Deep-learning research in educational data mining has therefore moved from final-score analysis to learning-process monitoring, behavior prediction, assessment, personalization and instructional support [1]. International research on artificial intelligence applications in higher education has formed four core directions: learner profiling and prediction, assessment and evaluation, adaptive learning systems, and intelligent tutoring systems [2]. Learner-generated materials, such as forum posts, reflective writings, course reviews and teacher feedback, also contain information that cannot be obtained through numerical scores; after appropriate preprocessing and interpretation, these texts can be used to support learning analytics [3].

The problem is not that the model has high scores on a single set of data; rather, it should be able to help students learn in real-life situations effectively. A recommendation system, an automatic grader and an AI teaching assistant process different data, produce different outputs, and thus have different levels of risk. CNNs, LSTMs and Transformers are compared to explore whether different architectures are more suitable for extracting local spatial or textual patterns (CNNs), retaining temporal dependencies (LSTMs), or capturing long-range context (Transformer-based models). Therefore, in selecting a model, one must consider prediction performance as well as inference speed, computational cost, privacy, fairness, explainability, factual accuracy and teacher supervision.

The organization will separate the mechanism, application evidence and decision criteria, and thus keep the recommendations transparent and educationally motivated. Section 2 presents the general ideas of the three architectures. Section 3 shows some typical cases of personalized learning, automated assessment and summarization, and knowledge-point question-answering. Section 4

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

compares the models according to educational scenarios and deployment requirements, and Section 5 summarizes the results and limitations.

2. Basic Principles of CNN, LSTM and Transformer

2.1. CNN

A convolutional neural network shares filters across the vicinity of an input and collects the recognized features by means of pooling or similar operations. Text: The input is a matrix of word or character embeddings, and filters act as detectors for informative phrases; Images: they recognize edges, shapes, handwriting, and other local structures. Kim demonstrated that a compact CNN with multiple filter widths and max pooling can effectively classify sentences [4]. Since the same filter is applied to all positions, convolution operations can be parallelized for high-speed inference. CNNs are thus more suitable for spatially local or concentrated evidence over a short range. The first is a short-term dependency constraint; a typical CNN cannot capture the relationship between distant words or parts of speech to understand the whole idea of a long text.

2.2. LSTM Network

A long short-term memory network processes a sequence step by step and keeps a hidden state and a memory cell. The input, forget and output gates control which information to add, keep or output, and reduce the loss of relevant signals that affect ordinary recurrent networks [5]. This structure is suitable for ordered events, such as a sequence of clicks, submissions, hints, dialogue turns or words; later predictions may depend on the preceding states. LSTM can model medium-term dependencies in time with relatively little data, and it can be extended bidirectionally to include both preceding and succeeding context. The first problem is that it is a sequential model; thus, the computation of a later state depends on its preceding state, and parallelization is not feasible.

2.3. Transformer

The Transformer replaces recurrence with self-attention. Each token assigns weights to other tokens and maintains the order of position; multiple attention heads can learn different relationships, such as references, contrasts, and topics [6]. Tokens in a layer can be processed in parallel, and pre-training can transfer contextual representations to other tasks. BERT, for example, learns bidirectional contextual representations and can be fine-tuned with relatively few labelled data [7]. Therefore, Transformer-based models are more suitable for long or semantically complex data. However, paying attention to long contexts and large pre-trained checkpoints can require considerable memory and computation. Pretraining also improves representation rather than ensuring factual correctness, so generation and high-risk educational use require additional controls.

2.4. Comparison

Figure 1 summarizes the different processing pipelines of the three architectures for educational data.

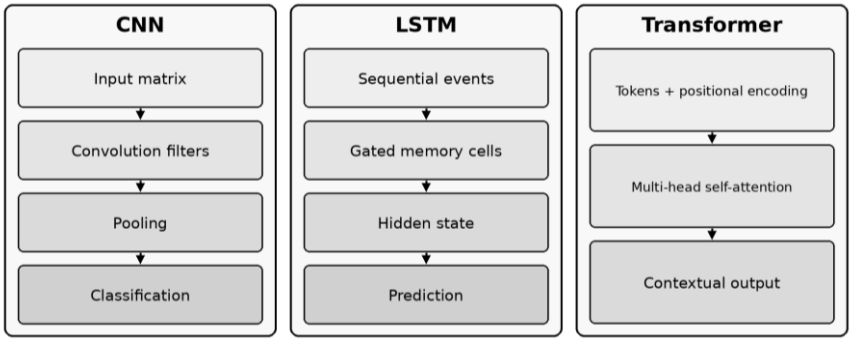

Figure 1. Processing Mechanisms of CNN, LSTM and Transformer for Educational Data (Picture credit: Original)

3. Representative Applications in Online Education

3.1. Personalised Learning and Dropout Prediction

A personalized learning system collects a student's learning history, behavior data and interests to recommend appropriate learning materials or adapt the difficulty of content. Research on emotion recognition of homework and forum messages shows that personalization can effectively link learner behaviors with their emotional expression [8]. Studies on MOOC discussion forums have also shown that learner discussions contain information on participation, course experience and sentiment that cannot be obtained from final Scores Alone [9].

Hybrid CNN-LSTM models are often employed in dropout prediction to learn jointly from local behavioral Features and their Temporal Evolution Patterns [10]. such combinations are suitable for conditions that need a short-term activity pattern and sequence order. It dropout prediction is only an early warning signal and does not indicate the final result judgment. missing events, uneven participation or platform-specific behavior patterns may lead to incorrect risk assessments. Therefore, prediction results should trigger teacher review and supportive intervention to limit a learner's learning automatically opportunities.

3.2. Automated Grading and Learning Summary

Recognize images, process natural language, automatically grade, generate homework problems and provide step-by-step solutions for the AI homework system. Tang et al. have built an intelligent homework ecosystem and coordinated the above functions to reduce teachers' repetitive grading and provide students with timely feedback. CNN components are naturally suitable for scanned handwriting, formulas, diagrams and other local visual patterns; sequential or contextual models can represent ordered solution steps and written answers.

Learning summarization extends to judging all responses. A platform can list completed activities, repeated misunderstandings, unresolved questions and recommended revisions from a study period. Transformer-based models are structurally suitable for semantic matching, long-input summarization and feedback generation, but the generated text may omit essential evidence or introduce unsubstantiated statements. The collected evidence has not provided a standardized comparison of the three architectures on the same grading dataset. Low-stakes practice can use automated feedback that shows uncertainty, and formal grades need explicit criteria, traceable evidence, and teacher supervision.

3.3. Knowledge-Point Question Answering and AI Tutoring

An AI teaching assistant identifies a learner's question, retrieves related course materials, maintains the dialogue context, and generates an explanation or follow-up resources. Huang and others built a course knowledge graph, added structured knowledge, and thus improved large-model question-answering performance; human evaluation also indicated enhancements in several areas of AI-related understanding [11]. Therefore, a good tutor needs to be able to present the learning material in a systematic way and also know when to present it.

A practical assistant should be based on a defined course repository, cite or present supporting materials, and distinguish between retrieved evidence and generated explanations in its answers. The controls reduce, but do not eliminate, reasonable but incorrect answers. AI tutors are suitable for the first clarification, navigation, practice questions and regular knowledge-point support. Serious academic disputes, private matters involving sensitive information and major behavioral deviations must be reported to the teacher; otherwise, as the school cannot investigate them, it is not responsible for the resulting harm.

3.4. Scenario-Based Comparison and Model Selection

Set the required input and output, latency, risk, etc., and then choose a model. Table 1 shows the general structural features; the following sections apply them to educational situations and only present data from similar study environments.

Table 1. Architectural Features and Scene Suitability

Model	Core mechanism	Suitable data	Educational scenarios	Main strength	Main limitation
CNN	Local convolution and pooling	Images, short text, local activity patterns	Handwriting and formula recognition; short-text classification; question routing	Fast parallel feature extraction; relatively compact	Limited long-range context
LSTM	Gated recurrent memory	Ordered text and behavior sequences	Learning-path and dropout prediction; dialogue or solution-step state	Preserves temporal order and sequential dependence	Stepwise computation limits parallelism
Transformer	Self-attention and positional information	Long or context-rich text; multi-source semantics	Open-ended grading; summarization; retrieval; AI tutoring	Strong contextual representation and transfer learning	Higher memory, serving, and governance requirements

4. Scenario-Based Comparison and Deployment Guidelines

4.1. Image and Short-Text Recognition Tasks

CNN is generally more suitable for handwriting, formula images, screenshots, short-text sentiment analysis and question-intent routing when the key evidence is local. Convolution can process positions in parallel, and compact models support high-throughput or low-latency services with limited hardware. LSTM is less effective without the order of events, and a large Transformer may incur extra serving costs. However, CNN should not be chosen solely for its speed when the decision requires consideration of distant qualifiers, subtle discourse or long-document meaning.

Koufakou directly compares the two on the same course-review corpus to show that a pretrained-word CNN achieved 92.1% accuracy and 82.4% Macro-F1, while RoBERTa reached 95.5% accuracy and 84.7% Macro-F1 [12]. As 91.5% of the reviews were positive, Macro-F1 is less sensitive to minority-class weaknesses. In the reported environment, the CNN trained in less than ten seconds per epoch, and the tested Transformer model took about three to four minutes. Select a case rather than an overall operating ratio.

4.2. Sequential Learner-Behaviour Modelling

LSTM is more suitable for ordered learner behavior that serves as evidence, such as click sequences, learning paths, repeated submissions, dialogue states and step-by-step problem-solving. Its gated memory retains the sense of time without the full expense of a large pre-trained language model. CNN can be added to LSTM to extract short local patterns, as shown in the dropout framework in [10], and Transformer may only be suitable for very long sequences or when substantial text needs to be integrated.

Intra-study comparisons are relatively strong evidence. Mrhar et al. compared CNN, LSTM, CNN-LSTM, and a Bayesian CNN-LSTM on MOOC sentiment-related data and reported the Bayesian hybrid as strongest across accuracy, precision, recall, and F1 [13]. Bai and others reported that a BERT-LSTM model increased the average F1 score by at least 6.19% for MOOC video prerequisite recognition across four subject areas, as shown by the authors [14]. These results support the effectiveness of hybrid representations for sequential tasks, but do not prove that a single architecture is universally optimal.

4.3. Long-Context Assessment, Summarisation and Tutoring

Transformer-based models generally have advantages in open-ended answer matching, long-text summarization, semantic retrieval and knowledge point tutoring, because these tasks require

understanding semantic relationships in large amounts of text. Pre-training also enables good transfer performance when labeled educational data is scarce. Related studies show that CNN-BiLSTM-attention models combining linguistic and sentiment knowledge can achieve good feedback analysis performance, but head-to-head comparisons on the same corpus still favor RoBERTa over pure CNN architectures [12]. Both results confirm the value of contextual representations, but do not prove that pure Transformer models are always superior.

Grading and tutoring tasks require compliance with scoring rubrics, factual correctness, retrieval relevance and expert judgment, which cannot be measured by a single accuracy metric. Therefore, Transformer generation systems should be integrated with scoring rubrics, course repositories, knowledge graphs or retrieval-augmented modules. Transformer models are cost-effective when the semantic performance gains cover the infrastructure and governance costs; for fixed, low-risk classification tasks that compact CNNs can handle well, large Transformer models are not an optimal choice.

4.4. Efficiency, Deployment Risk and Hybrid Models

Deployment complexity varies significantly across architectures. CNN generally has the lowest deployment complexity for stable local tasks. LSTMs require careful sequence construction, missing time step processing and sequence length tuning. Transformer systems may require larger checkpoint storage, retrieval services, prompt engineering, policy controls and monitoring systems, resulting in higher memory and operation costs. The carbon footprint and policy issues of large NLP models further increase the comprehensive cost beyond predictive indicators.

Educational risk factors can reverse purely technical rankings. Behavioral traces do not directly represent learning comprehension; sentiment analysis methods face challenges from ambiguity, code-switching and domain-specific terminology. A model performing well on one subject, institution or assessment system may fail to transfer to others. Research on the quality factors of courses also indicates that the essential-quality dimensions are not uniform. Therefore, all-encompassing model evaluation should also cover interpretability, uncertainty indication, access control, bias detection and appeal channels for outputs that affect grades or learning paths [15].

Most of the research mentioned above has used a hybrid model incorporating local, sequential and attention-based features. Hybrid architectures also have higher maintenance costs and are less interpretable. Therefore, the final selection criteria are conditional: CNN can be used for local pattern recognition in speed-priority scenarios and resource-constrained environments; LSTM is employed when the main feature is temporal order; and Transformer is used for long-context semantic analysis and generation under conditions of rich resources and governance. All high-stakes educational decisions involve human agency.

5. Conclusion

Systematically examine the application of CNN, LSTM and Transformer architectures in the core application scenarios of AI-enabled online education in this narrative review, and propose a scenario-adaptive model selection framework for educational practice. Research has formed three mature directions for the application of AI in online education so far: personalized learning intervention and dropout early warning, automated assessment support and learning summarization, and knowledge-point question answering and intelligent tutoring. The services handle various types of data, such as learner behavior logs, text submissions, image-based assignments and multimedia course materials, and thus impose different requirements for feature extraction and representation capabilities of deep learning models.

The buildings of these models are suitable for different purposes and have different applications. CNN has a good trade-off among inference speed, resource consumption and predictive performance, and is suitable for local visual or textual feature extraction, short-input classification tasks and deployment in resource-constrained environments. LSTM is still a good choice for learning path

modelling and medium-length behavior or language sequences where the primary evidence for prediction is temporal order. Transformer-based models generally have better performance in long-text understanding, complex semantic matching, open-ended assessment, content retrieval, automatic summarization and intelligent tutoring. However, these contexts-aware and generative capabilities are accompanied by higher infrastructure requirements, as well as an increased risk of factual hallucination and a heavier governance burden. A multi-level representation hybrid architecture can be used to improve the accuracy of prediction, but it is relatively complex to build and less intuitive.

Many non-technical reasons also affect how much actual educational value the deployed AI models have, such as data quality, protection of learner privacy, algorithmic fairness, model explainability, cross-scenario transferability, infrastructure conditions, and additional workload for teachers. As a narrative review, this study does not conduct controlled empirical experiments to uniformly compare the three architectures; the cited studies differ in task definition, dataset source, language environment, label distribution, hardware configuration and evaluation metric, so direct cross-study performance comparison is not feasible. Meanwhile, standardized comparative evidence on unified educational datasets remains insufficient, and the generalizability of conclusions across different teaching contexts needs further verification.

Future research will build a more all-encompassing evaluation system that includes not only predictive indicators, such as Macro-F1 and minority-class recall, but also deployment metrics, such as inference latency, memory consumption and energy consumption, as well as education-oriented attributes, including factual accuracy, explanation quality, algorithmic fairness, privacy protection and instructor workload. Organized comparative experiments on standardized educational datasets are required to develop more generalizable model selection criteria. Lightweight Transformer variants suitable for educational applications and interpretable hybrid architectures that balance performance and transparency also need to be explored further. Not all the above architectures are suitable for all educational environments in practice. The basic idea of model selection is to match the model's ability with the data structure of the platform, teaching goals, risk level and operating conditions. AI applications in education should always be aimed at improving the teaching capacity, and all high-stakes outputs, such as grades, learning interventions and academic advice, must be verified and finalized by human teachers.

References

[1] Z. Papamitsiou and A. A. Economides, Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence. Educ. Technol. Soc. **17**, 49–64 (2014)

[2] G.-J. Hwang and Y.-H. Tu, Artificial intelligence in education: a systematic review of research trends, applications, and challenges. Educ. Technol. Res. Dev. **71**, 1349–1382 (2023). https://doi.org/10.1007/s11423-022-10192-3

[3] N.-S. Chen, Y.-S. Chen, and S. J. H. Yang, Text mining in education: a systematic review of research trends, methodologies, and applications. Comput. Educ. **179**, 104421 (2022). https://doi.org/10.1016/j.compedu.2022.104421

[4] Y. Kim, Convolutional neural networks for sentence classification, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar (2014), 1746–1751. https://doi.org/10.3115/v1/D14-1181

[5] S. Hochreiter and J. Schmidhuber, Long short-term memory. Neural Comput. **9**, 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735

[6] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, in *Advances in Neural Information Processing Systems 30* (2017), 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

[7] J. Devlin, M.-W. Chang, K. Lee, et al., BERT: pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA (2019), 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] N. Altrabsheh, M. Cocea, and S. Fallahkhair, Sentiment analysis in MOOC forums: towards personalised learning support. Int. J. Artif. Intell. Educ. **31**, 517–545 (2021). https://doi.org/10.1007/s40593-020-00223-7

[9] W. Xing, X. Chen, J. Stein, et al., Mining MOOC discussion forums: a systematic review of research and applications. Internet High. Educ. **55**, 101862 (2022). https://doi.org/10.1016/j.iheduc.2022.101862

[10] J. Liang, D. Yang, Y. Wu, et al., MOOC dropout prediction using a CNN-LSTM hybrid model with behavioral features. Appl. Sci. **12**, 5678 (2022). https://doi.org/10.3390/app12115678

[11] X. Zhang, Y. Li, H. Wang, et al., Knowledge graph-based intelligent tutoring systems: a systematic review. Educ. Technol. Res. Dev. **72**, 589–623 (2024). https://doi.org/10.1007/s11423-023-10287-9

[12] A. Koufakou, Deep learning for opinion mining and topic classification of course reviews. Educ. Inf. Technol. **29**, 2973–2997 (2024). https://doi.org/10.1007/s10639-023-11736-2

[13] K. Mrhar, L. Benhiba, S. Bourekkache, et al., A Bayesian CNN-LSTM model for sentiment analysis in massive open online courses. Int. J. Emerg. Technol. Learn. **16**, 216–232 (2021). https://doi.org/10.3991/ijet.v16i23.24457

[14] Y. Sun, L. Wang, and Z. Wang, Prerequisite relation identification for MOOC videos using BERT-LSTM with semantic and structural features. IEEE Access **10**, 102345–102356 (2022). https://doi.org/10.1109/ACCESS.2022.3209876

[15] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, Explainable AI: a review of machine learning interpretability methods with educational applications. Entropy **23**, 1619 (2021). https://doi.org/10.3390/e23121619