Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: https://www.science-gate.com/IJAAS.html

Utilizing machine learning to align exam questions with program learning outcomes

Haifa Alharthi [1], *, Ashwaq Alhargan [1], Mohamed Habib [1, 2]

[1]*College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia*
[2]*Faculty of Engineering, Portsaid University, Port Said, Egypt*

ARTICLE INFO

Article history:
Received 3 August 2025
Received in revised form
20 December 2025
Accepted 3 January 2026

Keywords:
Machine learning
Program learning outcomes
Exam question mapping
Educational assessment
Quality assurance

ABSTRACT

Aligning exam questions with course learning outcomes and linking them to program learning outcomes is often a time-consuming process that is prone to human error. This study examines the effectiveness of machine learning techniques for automatically mapping exam questions to Program Learning Outcomes (PLOs) and performance levels. A dataset of 414 multiple-choice questions was used to develop prediction models based on both joint and single-model architectures. The results show that the automated models achieved higher accuracy than human evaluations, indicating strong potential for the use of AI-based tools in educational quality assurance. The proposed approach can support academic institutions by automating assessment-related tasks, reducing faculty workload, and improving curriculum alignment. To the best of our knowledge, this study is the first to address the automated mapping of exam questions to program learning outcomes using machine learning methods.

1. Introduction

In recent decades, higher education has faced pressure to improve academic quality because of concerns about slipping standards, a more globalized landscape, and the need for efficient learning methods. Quality assurance is a key response to these challenges (Aamodt et al., 2018). Multiple international organizations have provided standards for program accreditation to ensure the quality assurance of different academic programs. This includes organizations such as the Accreditation Board for Engineering and Technology (ABET), which specializes in accrediting programs in science, technology, engineering, and mathematics (STEM), and the Southern Association of Colleges and Schools Commission on Colleges (SACSCOC), which grants accreditation to higher education institutions.

In addition, many countries have started national accreditation organizations to set standards for academic programs. In the Kingdom of Saudi Arabia, the Higher Education Council established the National Commission for Academic Accreditation and Assessment (NCAAA). The NCAAA acts as a gatekeeper of educational quality in Saudi Arabia. Their tasks encompass accrediting both institutions and individual programs, overseeing institutions seeking international recognition, and monitoring the ongoing quality of the accredited programs. Additionally, the NCAAA conducts evaluations of domestic institutions and programs and collaborates with relevant organizations both within and outside the kingdom.

Accreditation emphasizes an institution's ability to achieve its goals. One of the key goals of colleges is student learning. Accreditors assess this by considering student outcomes, the college's mission, learning objectives, and how they measure student progress. The challenge for colleges lies in defining their specific student learning goals, considering their mission and curriculum, and creating ways to measure them effectively (Beno, 2004). To achieve accreditation, institutions must demonstrate that they effectively equip students with the knowledge and skills necessary for success. Program learning outcomes (PLOs) play a critical role in this process. PLOs are clearly defined statements that outline the specific competencies students will acquire at the end of the program. They define the knowledge, skills, and abilities that students are expected to master by the end of a program. Accrediting agencies evaluate how well the curriculum and assessment methods of a program align with PLOs. This ensures that the program delivers on its promises, producing

* Corresponding Author.
Email Address: h.alharthi@seu.edu.sa (H. Alharthi)
https://doi.org/10.21833/ijaas.2026.01.015
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-9267-8497

graduates who are well-prepared for their chosen field. Accreditation acts as a public seal of approval, whereas PLOs provide a roadmap for achieving this distinction. Table 1 presents three examples of program learning outcomes categorized into the domains of Knowledge and Understanding, Skills, and Values, Autonomy, and Responsibility.

Table 1: Domains of learning and their corresponding program learning outcomes

Domains of learning	PLO statement
Knowledge and understanding	Recognize the major theories of machine learning techniques including neural networks.
Skills	Explore, analyze, manage, and visualize large data sets using the latest technologies.
Values, autonomy, and responsibility	Evaluate opportunities to employ data science solutions in accordance with business ethics and values.

Program Learning Outcomes serve as general guidelines for structuring an academic program, while course learning outcomes (CLOs) are defined with greater specificity, outlining the knowledge, skills, or values that students are expected to attain at the end of a particular course. These outcomes directly align with the course content and activities while contributing to broader PLOs.

Course performance levels are denoted by I (Introduction), P (Practiced), and M (Mastered) to signify the stages in which learning outcomes are integrated into a program's curriculum. "I" (Introduction) indicates that students are introduced to the PLO, establishing the essential knowledge and comprehension required for further learning. The level "P" (Practiced) means that the PLO is strengthened through experiential activities, including tutorials, labs, or discussions centered on case studies, thereby allowing students to enhance their skills. The level "M" (Mastered) indicates that students have achieved adequate practice to demonstrate mastery over the skills or knowledge connected to the PLO.

Strong academic programs are built on the foundation of program learning outcomes. They serve as a roadmap, explicitly defining the competencies that students will have at graduation. Having well-defined PLOs benefits universities, colleges, academic staff, and students. After gaining a thorough understanding of the program's expectations, staff members can design courses that precisely address these outcomes, ensuring a consistent curriculum. PLOs offer a structure for academic program evaluation that enables organizations to measure student progress and consistently improve program efficiency.

Beyond their internal benefits, program learning outcomes play a critical role in attaining program accreditation. Accreditation from a recognized organization such as ABET indicates that a program meets conventional quality standards. Accrediting organizations use PLOs to evaluate how efficiently a program prepares students for their major. By demonstrating the alignment between PLOs, curricula, and assessment methods, institutions can show accreditors that their program equips graduates with the necessary knowledge and skills. This improves a program's reputation and appeal to prospective students in addition to securing accreditation. Explicit Program Learning Outcomes provide students with a clear understanding of academic expectations and methods for self-evaluation while also offering employers and accrediting bodies proof of the program's success in imparting the necessary knowledge and skills to graduates (Japee and Oza, 2021).

Program learning outcomes act as powerful tools for strengthening assessment practices. By aligning assessments with PLOs, educators ensure that they are measuring what truly matters: student achievement of core program objectives. This targeted approach avoids irrelevant assessments and provides valuable data for improvement. Effective assessment strategies have been applied, for example, direct assessments, which measure PLO achievement directly (e.g., exams), with indirect assessments that assess broader skills (e.g., surveys and interviews), providing a holistic assessment of student learning outcomes. Additionally, crafting assessments that mirror real-world scenarios allows students to apply their knowledge and skills in practical contexts, thereby authentically demonstrating their mastery of PLOs (Gao et al., 2020). Thus, well-defined program learning outcomes pave the way for effective assessment.

In assessing student learning, it is crucial to ensure that assessments align with their intended purposes. This can be accomplished by linking each assessment question to specific Course Learning Outcomes (CLOs). By mapping questions with individual CLOs, educators can verify that the assessments measure the correct learning objectives of the course and that the questions thoroughly cover all the course goals. CLOs are designed to contribute to broader PLOs, which represent the principal goals of the program. This hierarchical mapping helps educators track how individual assessment items support the achievement of program-level goals.

In addition, it provided students with a clear understanding of the objective of each question and its relevance to their learning progression. This systematic approach also facilitates the collection of valuable data, enabling educators to analyze the extent to which students achieve specific CLOs and PLOs. Such insights can guide continuous improvement efforts, ensuring that assessments, courses, and programs remain aligned with the intended outcomes and meet the accreditation requirements.

Manually linking questions in assignments and exams to CLOs and subsequently to PLOs is time-consuming and error-prone. This task required instructors to carefully assess each question and align it with the relevant CLO. The process becomes even more challenging in programs with large question banks, a large number of learning outcomes, and interdisciplinary courses. Inconsistencies and personal interpretations can arise, leading to potential misalignments between assessments and learning objectives. This challenge can be effectively addressed using machine learning (ML) techniques.

This research tackles a crucial challenge in education: ensuring that assessments accurately measure students' achievement of program learning outcomes. We present a novel approach that automatically predicts which PLO a given assessment question belongs to. This can significantly increase the effectiveness and efficiency of the assessment design. By leveraging natural language processing (NLP), ML models can analyze the textual content of questions and map them to predefined PLOs. Teachers can increase the regularity and quality of assessment alignment with program goals, while also saving a significant amount of time by automating this process. This ultimately leads to a more robust assessment system that truly reflects student learning in the program.

We examined the effectiveness of employing machine learning to link exam questions with program learning outcomes and performance levels. A dataset of 414 multiple-choice questions from both midterm and final exams within the data science curriculum was gathered. Various machine learning and deep learning techniques have been applied to predict PLOs and their performance levels (I, P, and M), both jointly and separately. The best-performing model predictions were analyzed using human evaluations. To the best of our knowledge, this is the first study to automate the process of exam question mapping to program learning outcomes.

Given that CLOs are specific to individual courses, accurate prediction requires a large number of questions per course. However, given that every course is designed with specific CLOs that align with the PLOs, mapping them is straightforward. In cases in which CLOs are uniquely linked to individual PLOs, determining the CLO associated with a question becomes straightforward once the predictive model identifies the corresponding PLO. However, in scenarios where multiple CLOs are mapped to the same PLO, the mapping process becomes more complex. In such instances, the model can suggest potential CLOs for the instructor to choose, thus enabling informed decisions based on the context of the question. This approach not only simplifies the mapping process but also provides flexibility for instructors to validate and refine the mappings as needed.

The remainder of this paper is organized as follows. Section 2 examines the relevant literature. Section 3 describes the proposed methodology, and Section 4 details the evaluation approach, including the materials used and experimental settings. The results and a discussion of the findings are presented in Section 5. Finally, Section 6 summarizes the key findings, highlights their contributions to the field, and discusses future research.

2. Literature review

Natural language processing and machine learning have been increasingly utilized in quality assurance and various educational processes, ranging from automated assessment grading (Valenti et al., 2003) to curriculum alignment (Pattnaik et al., 2024) and personalized learning (Mathew et al., 2021).

Ujkani et al. (2021) proposed a system to analyze learning outcomes from syllabi and program curricula by identifying inconsistencies. This ensures that the programs deliver the knowledge and skills promised to students. By aiding both quality assurance officers and lecturers, the system aims to contribute to a more robust quality assurance process in universities. Putri et al. (2022) presented an overview of how artificial intelligence (AI), machine learning, and deep learning (DL) transform education for students, educators, and administrators. This study proposes a new way to examine the role of AI across the entire educational journey. This framework considers proactive planning (admission and course scheduling) and reactive execution (knowledge delivery and assessment). This review analyzes 194 research articles published between 2003 and 2022 to identify key research trends in AI-driven education for both the proactive and reactive phases. It explores the evolution of the choice of data and algorithms used in the AI solutions over time. This review also examines the impact of the COVID-19 pandemic on education and how it accelerates the adoption of AI tools. Finally, it discusses the limitations of AI in education and suggests directions for future research and development.

Transfer learning techniques were adopted by Lagus et al. (2018) to boost the prediction accuracy of learning outcomes, particularly when dealing with limited training data, such as when making early predictions in a new setting. However, improvements in predictive power are often modest. Traditional machine learning models can still be quite accurate, as long as the contexts being compared are very similar and the student activity features are designed to minimize the influence of minor differences between those contexts. Zaki et al. (2023) introduced an AI system that automates the mapping of learning outcomes between courses and programs. The system uses natural language processing to analyze the text and automatically perform mapping. Tests using real data from two educational programs showed promising results. The AI system achieved high accuracy (over 83%) compared to human experts performing the same task. These findings suggest that the proposed AI framework has significant potential to streamline this process. Shaikh et al. (2021) proposed a new method for classifying learning objectives and assessments according to Bloom's taxonomy, a framework that categorizes educational goals based on cognitive complexity. Current methods using keywords have low accuracy. Shaikh et al. (2021)

addressed this by proposing a deep learning model using Long Short-Term Memory (LSTM) networks, which achieved significantly higher accuracy (87% for learning objectives and 74% for assessments) than keyword-based approaches (55% accuracy). The simplicity of the model makes it appealing, and its performance surpasses that of previous attempts at this task. In Supraja et al. (2017), a new system was introduced that automatically matched assessment questions with learning goals. The authors used a simplified version of a common framework for classifying learning objectives (Bloom's taxonomy). Their research reduced Bloom's taxonomy into three categories: remember (involving lower-order thinking), apply (focused on application-based questions), and transfer (involving higher-order thinking that necessitates analysis and synthesis). The system analyzed the text of the questions and assigned labels based on the intended learning goals. It uses techniques to convert the question text into a format that the system can understand. They trained it on questions labeled by an expert, and then validated its performance using real-world questions from online sources. The results showed that their system performed very well (86% accuracy) compared to a human expert, indicating its potential to improve the efficiency and effectiveness of assessments in education.

3. Methodology

This study examined two strategies for predicting program learning outcomes and performance levels: the joint model and the single model.

3.1. Joint model approach

This approach uses a single neural framework to simultaneously predict both labels. The architecture of the joint model consists of the components shown in Fig. 1.

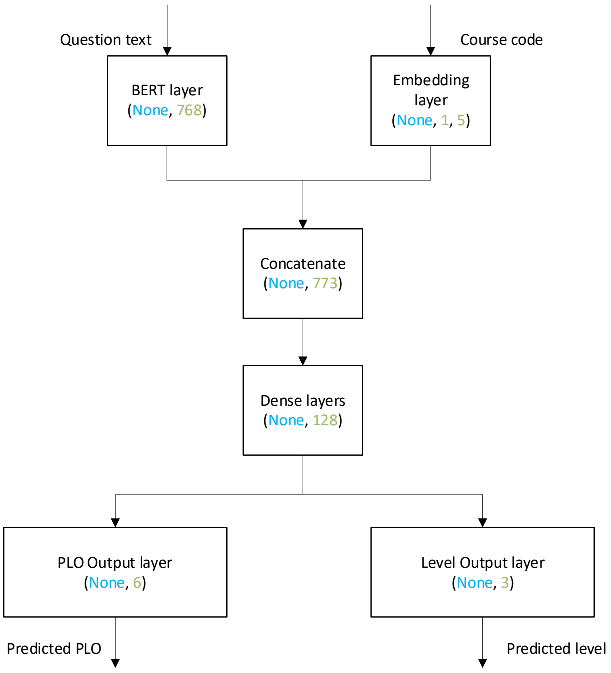


Fig. 1: The proposed joint model architecture

Question input layer: This layer takes a sequence of tokens and encodes them using a pre-trained BERT model, particularly the bert-base-uncased checkpoint from the HuggingFace Transformers library. The output is a dense representation acquired by mean-pooling the final hidden states, which results in a 768-dimensional embedding that captures the semantics of the textual input.

Course code layer: This embedding layer transforms each of the 11 course codes into a trainable 5-dimensional dense representation. Subsequently, a flattened layer is applied to reshape these vectors into a one-dimensional array.

Concatenate layer: This layer combines the outputs from the preceding layers, namely BERT-encoded questions and dense course code representations, and creates an integrated representation vector consisting of 773 dimensions.

Fully connected layers: 0, 1, or 2 fully connected layers are added after the concatenation layer. We experimented with multiple numbers of hidden layers and different neuron sizes.

PLO output layer: A softmax output layer with six units representing the number of PLO classes.

Level output layer: A softmax output layer with three units representing the number of performance levels.

The model was compiled using the Adam optimizer. Because both outputs require multiclass predictions, the categorical cross-entropy function was employed as the loss function, and accuracy was used as the evaluation metric for each output.

3.2. Single model

Here, the multilabel classification task is decomposed into two independent classification tasks, in which each label is predicted separately. Unlike the joint model, this approach assumes a lack of dependence between labels. We considered multiple classifiers: Logistic regression (LR), support vector machine (SVM), and random forest (RF). The PLO and performance-level classifiers were trained on the same input features; however, they were independently optimized. Three feature extraction methods were used: TF-IDF vectorization, GloVe, embeddings (Pennington et al., 2014), and BERT embeddings (Devlin et al., 2019).

4. Evaluation

In this section, we describe the evaluation framework in detail. Fig. 2 shows an overview of the framework, comprising the collection of the dataset, preprocessing and extraction of features, and development and evaluation of models.

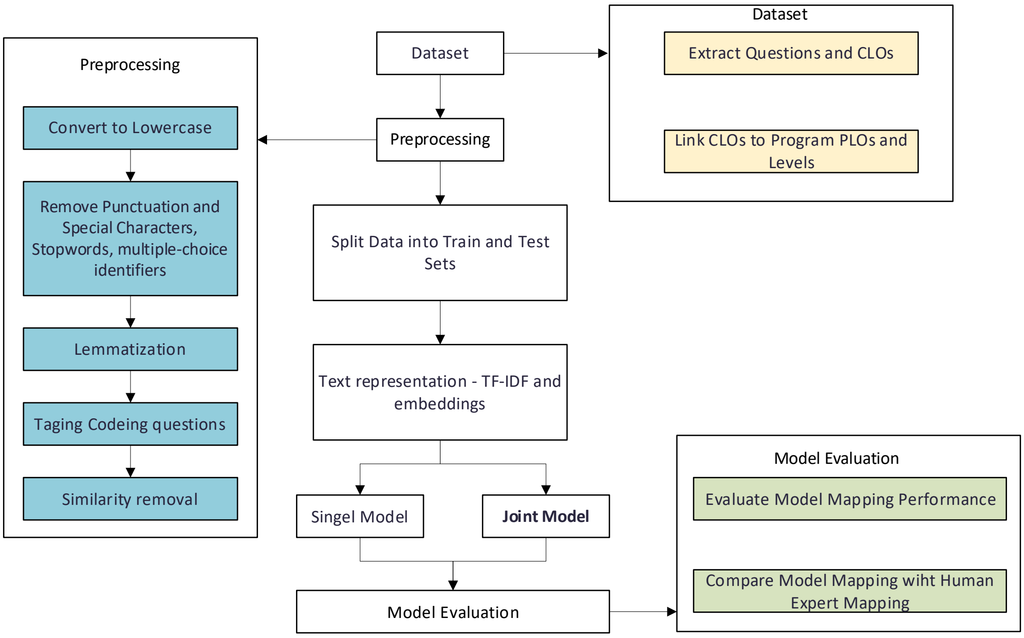

Preprocessing
Convert to Lowercase
Remove Punctuation and Special Characters, Stopwords, multiple-choice identifiers
Lemmatization
Taging Codeing questions
Similarity removal
Dataset
Preprocessing
Split Data into Train and Test Sets
Text representation - TF-IDF and embeddings
Singel Model
Joint Model
Model Evaluation
Dataset
Extract Questions and CLOs
Link CLOs to Program PLOs and Levels
Model Evaluation
Evaluate Model Mapping Performance
Compare Model Mapping wiht Human Expert Mapping

Fig. 2: Overview of the proposed framework

4.1. Dataset

We collected 414 English multiple-choice exam questions from 11 courses of the Data Science Master's Program. Each question is linked to a Course Learning Outcome (CLO) that is associated with a program learning outcome (PLO) and performance level. An example question, with its corresponding CLO, PLO, and Level, is presented in Table 2. Given that CLOs are specific to individual courses, accurate prediction requires a large number of questions per course. Our main objective, however, is to determine the PLO and Level for each question. There are six PLOs in the Data Science Master's program. Level, however, categorizes performance into three distinct classes: I, P, and M.

4.2. Metrics

We evaluated performance using accuracy, weighted precision, recall, and F1 score. These weighted metrics consider the class imbalance by considering w_i, which denotes the proportion of true instances for each class i among the total number of classes C, calculated as follows:

$$w_i = \frac{\text{Number of true instances in class } i}{\text{Total instances}}$$

$$\text{Weighted Precision} = \sum_{i=1}^{C} w_i \cdot \text{Precision}_i$$

$$\text{Weighted Recall} = \sum_{i=1}^{C} w_i \cdot \text{Recall}_i$$

$$\text{Weighted F-score} = 2 \cdot \frac{\text{Weighted Precision} \cdot \text{Weighted Recall}}{\text{Weighted Precision} + \text{Weighted Recall}}$$

4.3. Experimental settings

4.3.1. Preprocessing

To prepare the text for analysis, we implemented standard preprocessing techniques. We converted the question texts to lowercase, removed punctuation words, and eliminated the NLTK stop words. The NLTK English stop word list includes WH question words (e.g., what and where), which may be important for the classification of questions. Consequently, these terms were excluded from the list of stop words before their removal. Furthermore, we deleted multiple-choice identifiers such as "a." and "b." which are part of any multiple-choice question. Subsequently, we conducted lemmatization to convert the words into their root forms using WordNet.

Table 2: An example of one question with its associated CLO, PLO, and level

Course	Question	CLO	PLO	Level
DS510	What will be the probability of getting odd numbers if a dice is thrown? a. 1/2 b. 2 c. 4/2 d. 5/2	LO1	K2	I

Many questions contain code snippets that may be unrecognized by pre-trained word-embedding models; hence, we enhanced these questions with related terms such as” programming” and “code.” For this purpose, we utilize regular expressions to identify patterns resembling code such as “def” and “const.” To identify near-duplicate questions that were slightly paraphrased, we calculated cosine similarities between the question pairs. If the similarity exceeded the threshold of 80%, we discarded a similar question.

4.3.2. Single model settings

The dataset was divided into training and testing sets in an 80:20 ratio. The distribution of PLO labels in the training set indicated a significant imbalance. Classes V1 and S1 constituted approximately 4% and 6% of the data, respectively, while S2 and K2 accounted for approximately 34% and 27%. To address this disparity, we employed random oversampling to increase the representation of underrepresented PLO classes, while maintaining the natural distribution of levels. Oversampling was exclusively applied to the training set.

We used sklearn to conduct the experiments and fine-tuned the logistic regression, support vector machines, and random forest. To determine the best hyperparameters, we performed a grid search cross-validation approach with imbpipeline (imbalanced-learn.org), a pipeline used to split the training set into five segments and apply random oversampling to the training folds. The classifiers were trained using oversampled segments and evaluated using the validation segment. Table 3 lists the hyperparameters evaluated for each algorithm and the best-performing algorithms.

A sklearn column transformer was used to independently transform each column. The TF-IDF weights are learned from the textual input of the oversampled training segment, and the learned transformation is applied to the validation set. The course input was transformed into one-hot encoding, and the PLO and Level labels were converted into numerical values. We used fifty-dimensional pre-trained GloVe word vectors to represent words and then compute their average. Similarly, we used BERT embeddings to encode the words and then averaged them.

4.3.3. Joint model settings

The same training and testing splits were maintained, and a random validation split of 20% was adopted. The texts were tokenized and represented using the BERT model (Devlin et al., 2019). Embeddings were used to represent the categorical course inputs. Both output layers used the categorical cross-entropy loss function during training. Early stopping was adopted to observe the total validation loss and to stop the training process if there was no improvement for three epochs (Fig. 3). We experimented with multiple hyperparameters and selected those that yielded the best performance in the validation set (Table 3). The hyperparameters are embedding sizes of 5 and 10, dropout rates of 0, 2, and 4, and the number of hidden layers post-concatenation of 0, 1, and 2.

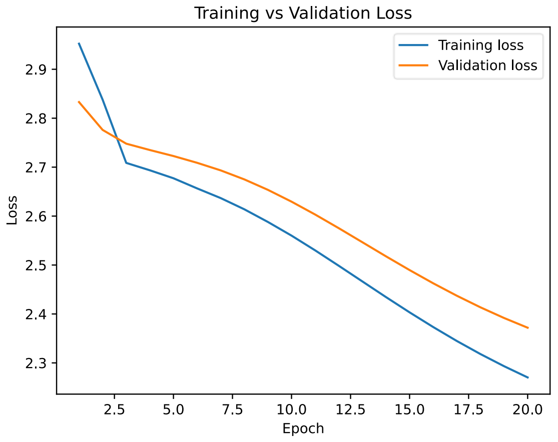


Fig. 3: Training and validation loss during training the joint model

Table 3: Best hyperparameters

Model	Oversampling	Best hyperparameters
Joint model	No	number of hidden layers=0 embedding dim=5; drop rate=0; epochs=20; batch size=16
BERT-LR	No	C: 1, penalty: l2
BERT-SVC	No	C: 10, kernel: rbf
GLOVE-SVC	No	C: 1, kernel: poly
GLOVE-LR	No	C: 1, penalty: l2
TF-IDF-SVC	No	C: 10, kernel: rbf
TF-IDF-SVC	No	C: 10, kernel: poly
BERT-SVC	Yes	C: 1, kernel: poly
BERT-SVC	Yes	C: 1, kernel: linear
GLOVE-SVC	Yes	C: 10, kernel: poly
GLOVE-SVC	Yes	C: 1, kernel: linear
TF-IDF-SVC	Yes	C: 1, kernel: rbf
TF-IDF-LR	Yes	C: 10, penalty: l2

4.3.4. Human-evaluation

Because mapping the exam questions to the corresponding learning outcomes is subjective and dependent on the instructor’s personal interpretations, we compared the model predictions to the judgments of human evaluators. Three independent faculty members with domain experience who had taught more than two courses in the program were recruited to manually map the questions in the test set to their corresponding CLOs, PLOs, and performance levels. Each instructor received a list of test set questions along with the corresponding course code, as shown in Table 2, and was asked to select the learning outcomes that best aligned with each question based on their teaching experience in the program.

Inter-rater reliability: Fleiss' Kappa statistical analysis (Fleiss, 1971) was applied to measure the degree of agreement among the three faculty members on their mapping of the questions to PLOs and Levels. CLOs mapping was excluded from the analysis as they are specific to individual courses and are associated with a program, PLOs, and performance levels. The results showed fair agreement between faculty members, with k = 0.341 and k = 0.38, for PLOs and level mapping, respectively. This result indicates that faculty might have different perspectives or interpretations of the questions, which makes reliable mapping of the questions a challenging task.

5. Results and discussion

We conducted cross-validation for the combination of single models (i.e., SVC, LR, and RF) and data representations (i.e., BERT, GloVe, and TF-IDF), and whether random oversampling was applied. In addition to the joint classifier results, Table 4 presents a comparison of the best "Performance Level" classifiers for each of the data representations, both with and without oversampling. Fig. 4 shows the evaluation of the prediction accuracy across the validation and test sets. The TF-IDF-SVC model achieved the highest performance without oversampling, yielding a test set result of 0.68 accuracy. The impact of oversampling varied, producing classification accuracies that were 15% and 8% lower for BERTSVC and TF-IDF-LR, respectively, and a 2% increase for GLOVE-SVC.

Cross-validation was similarly performed to identify the best combination of PLO single machine-learning classifiers with data representations, including the application of random oversampling. Fig. 5 shows the validation and test classification accuracies achieved by the top models. Table 5 presents the test results for the joint and single models.

Overall, for all classifiers and evaluation metrics, the results of the PLO classification were at least 13% lower than the "performance level" predictions. This difference is expected because PLO involves six labels, whereas the levels have only three labels. Furthermore, the TF-IDF-SVC approach, without the use of oversampling, delivered optimal performance in every metric with an accuracy of 0.55% and an F-score of 0.51%. Random sampling continues to yield inconsistent outcomes, showing increased f-scores for BERT-SVC and GLOVE-SVC but reduced accuracy and f-score for TF-IDF-SVC.

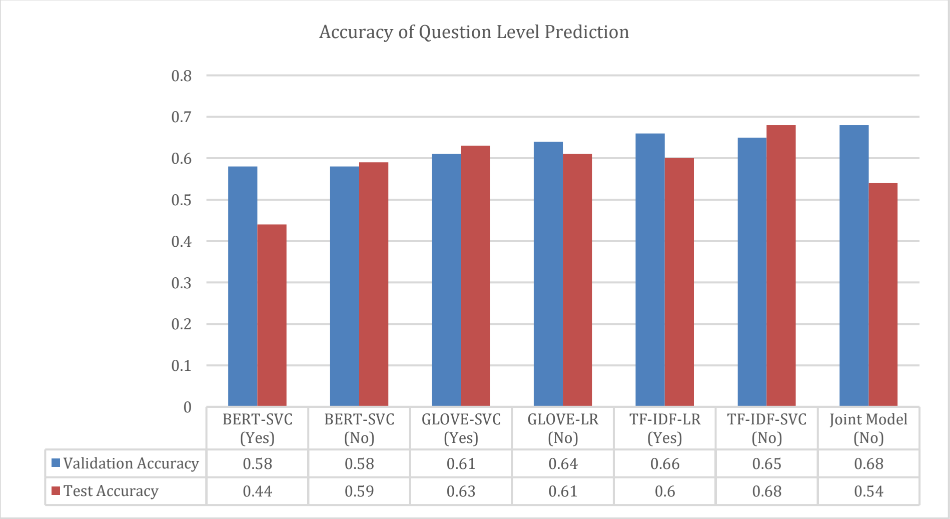


Fig. 4: Validation and test accuracy for question-level prediction

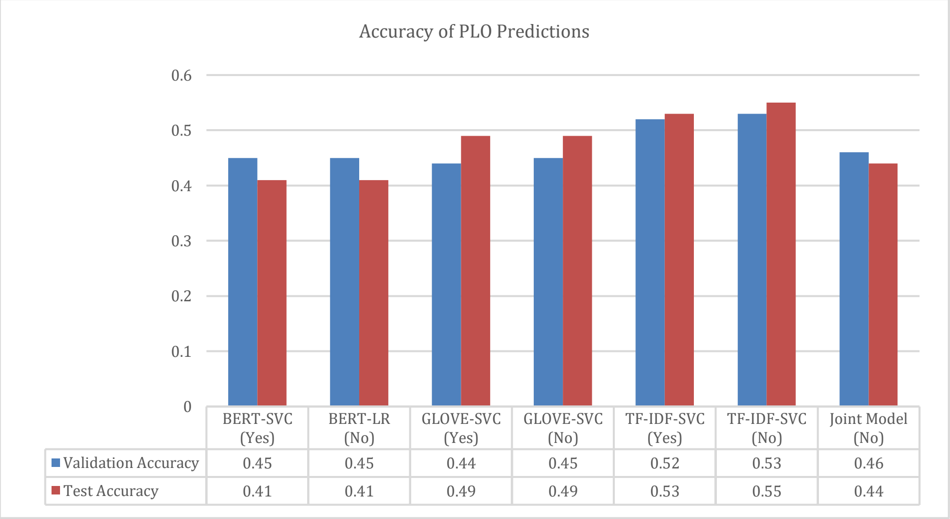
Table 4: Test results for question performance level prediction

Model	Oversampling	Accuracy	Precision	Recall	F1-score
BERT-SVC	Yes	0.44	0.45	0.44	0.43
	No	0.59	0.60	0.59	0.58
GLOVE-SVC	Yes	0.63	0.63	0.63	0.62
	No	0.61	0.61	0.61	0.61
TF-IDF-LR	Yes	0.60	0.60	0.60	0.60
TF-IDF-SVC	No	0.68	0.68	0.68	0.68
Joint model	No	0.54	0.54	0.54	0.53

The superior performance of the TF-IDF and SVM models compared to that of the BERT-based model could be attributed to the small size of the training data. Support Vector Machines are recognized for requiring only a few instances to identify the maximum-margin hyperplane that is essential for classification tasks (Moguerza and Muñoz, 2006). However, fine-tuning a BERT model requires a large task-specific dataset to effectively modify its parameters. Therefore, it is recommended that this dataset be expanded in future studies to evaluate this hypothesis. To further understand the performance of TF-IDF-SVC without the use of oversampling, Tables 6 and 7 show a breakdown of the precision, recall, and f-score for each PLO and performance-level label. The model performed well for some PLO classes (e.g., Class 3) but struggled with others, especially those with low support (Classes 2 and 5). All three level classes have similar F1-scores (0.65–0.70), indicating fairly consistent performance across the board.

Table 5: Test results for PLO prediction

Data representation	Oversampling	Accuracy	Precision	Recall	F1-score
BERT-SVC	Yes	0.41	0.37	0.41	0.39
BERT-LR	No	0.41	0.38	0.41	0.37
GLOVE-SVC	Yes	0.49	0.46	0.49	0.47
GLOVE-SVC	No	0.49	0.41	0.49	0.41
TF-IDF-SVC	Yes	0.53	0.48	0.53	0.50
TF-IDF-SVC	No	0.55	0.50	0.55	0.51
Joint model	No	0.44	0.30	0.44	0.35



	BERT-SVC (Yes)	BERT-LR (No)	GLOVE-SVC (Yes)	GLOVE-SVC (No)	TF-IDF-SVC (Yes)	TF-IDF-SVC (No)	Joint Model (No)
🟦 Validation Accuracy	0.45	0.45	0.44	0.45	0.52	0.53	0.46
🟥 Test Accuracy	0.41	0.41	0.49	0.49	0.53	0.55	0.44

Fig. 5: Cross-validation and test accuracy for PLO prediction

The performance of faculty members' PLOs and level mappings was compared with the ground truth labels of the test set (the ground truth refers to the given labels in the collected dataset). As shown in Table 8, the results illustrate that on average, faculty members performed at an accuracy level of 0.42 and 0.57 for PLO and Level mapping, respectively, with moderate precision and recall scores. It can be seen from the result that there are differences in the faculty performance, which highlights the variability in how they interpreted the mapping task. Interestingly, faculty members 2 and 3 demonstrated higher accuracy than faculty member 1, which could be attributed to their extensive teaching experience within the program.

Table 6: Classification report for PLO predictions

Question's PLO label	Precision	Recall	F1-score	Support
0	0.38	0.33	0.35	9
1	0.48	0.74	0.58	19
2	0.00	0.00	0.00	7
3	0.69	0.79	0.73	28
4	0.63	0.38	0.48	13
5	0.00	0.00	0.00	4

Table 7: Classification report for performance level predictions

Question's Level	Precision	Recall	F1-score	Support
0	0.63	0.76	0.69	25
1	0.79	0.63	0.70	24
2	0.65	0.65	0.65	31

Table 8: Faculty members' mapping results compared with the ground truth labels

Faculty	Accuracy	Precision	Recall	F1-Score
		PLOs		
1	0.33	0.36	0.33	0.34
2	0.48	0.49	0.48	0.48
3	0.45	0.49	0.45	0.44
Average	0.42	0.45	0.42	0.42
		Levels		
1	0.53	0.53	0.53	0.52
2	0.60	0.64	0.60	0.60
3	0.60	0.60	0.60	0.60
Average	0.57	0.59	0.57	0.57

These findings indicate that question-mapping is difficult for faculty members. Improving mapping guidelines and providing additional training could lead to more accurate and consistent results for faculty members. Comparing the human evaluation results in Table 8 with the best model predictions (Figs. 3 and 4), it is clear that the model predictions outperform human predictions by 12% and 10% for PLOs and Levels, respectively. This suggests that, while the task is quite difficult for humans, a machine learning model might offer a more reliable and efficient solution, highlighting the need to automate the question-mapping task. The superior performance of ML models compared to that of human experts implies that most labels adhere to a consistent pattern. The model is likely to be identified and generalized from this dominant pattern, possibly ignoring anomalies and noise. The model may have accurately captured the decision-making processes used by course instructors when labelling questions.

6. Conclusion

This study examines the effectiveness of machine learning techniques for mapping exam questions to program learning outcomes and associated performance levels. We evaluated multiple machine learning algorithms across three different feature representations alongside a deep learning–based joint model. To ensure robust evaluation, we performed extensive cross-validation experiments to optimize the hyperparameters both with and without oversampling. The Support Vector Machine using TF-IDF features without oversampling achieved the highest accuracy in predicting both PLOs and performance levels. We then compared the model's predictions with those of human evaluators and found that the model consistently outperformed human judgments, highlighting the significant potential of AI-powered tools for enhancing quality assurance in educational systems.

List of abbreviations

ABET	Accreditation Board for Engineering and Technology
AI	Artificial intelligence
BERT	Bidirectional Encoder Representations from Transformers
CLO	Course learning outcome
CLOs	Course learning outcomes
DL	Deep learning
EMNLP	Empirical Methods in Natural Language Processing
GloVe	Global Vectors for word representation
LR	Logistic regression
LSTM	Long short-term memory
ML	Machine learning
NCAAA	National Commission for Academic Accreditation and Assessment
NLP	Natural language processing
PLO	Program learning outcome
RF	Random forest
SACSCOC	Southern Association of Colleges and Schools Commission on Colleges
STEM	Science, technology, engineering, and mathematics
SVM	Support vector machine
SVC	Support vector classifier
TF-IDF	Term frequency–inverse document frequency

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Aamodt PO, Frølich N, and Stensaker B (2018). Learning outcomes: A useful tool in quality assurance? Views from academic staff. Studies in Higher Education, 43(4): 614-624. https://doi.org/10.1080/03075079.2016.1185776

Beno BA (2004). The role of student learning outcomes in accreditation quality review. New Directions for Community Colleges, 2004(126): 65-72. https://doi.org/10.1002/cc.155

Devlin J, Chang MW, Lee K, and Toutanova K (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Minneapolis, USA, 1: 4171-4186. https://doi.org/10.18653/v1/N19-1423

Fleiss JL (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5): 378-382. https://doi.org/10.1037/h0031619

Gao X, Li P, Shen J, and Sun H (2020). Reviewing assessment of student learning in interdisciplinary STEM education. International Journal of STEM Education, 7: 24. https://doi.org/10.1186/s40594-020-00225-4

Japee G and Oza P (2021). Curriculum and evaluation in outcome-based education. Psychology and Education Journal, 58(2): 5620-5625. https://doi.org/10.17762/pae.v58i2.2982

Lagus J, Longi K, Klami A, and Hellas A (2018). Transfer-learning methods in programming course outcome prediction. ACM Transactions on Computing Education, 18(4): 19. https://doi.org/10.1145/3152714

Mathew AN, Rohini V, and Paulose J (2021). NLP-based personal learning assistant for school education. International Journal of Electrical and Computer Engineering (IJECE), 11(5): 4522-4530. https://doi.org/10.11591/ijece.v11i5.pp4522-4530

Moguerza JM and Muñoz A (2006). Support vector machines with applications. Statistical Science, 21(3): 322–336. https://doi.org/10.1214/088342306000000493

Pattnaik P, Maheshwary R, Ogueji K, Yadav V, and Madhusudhan ST (2024). Enhancing alignment using curriculum learning and ranked preferences. In: Al-Onaizan Y, Bansal M, and Chen YN (Eds.), Findings of the association for computational linguistics: EMNLP 2024: 12891-12907. Association for Computational Linguistics, Miami, USA. https://doi.org/10.18653/v1/2024.findings-emnlp.754

Pennington J, Socher R, and Manning CD (2014). GloVe: Global vectors for word representation. In the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar: 1532-1543. https://doi.org/10.3115/v1/D14-1162

Putri NSF, Widiharso P, Utama ABP, Shakti MC, and Ghosh U (2022). Natural language processing in higher education. Bulletin of Social Informatics Theory and Application, 6(1): 90-101. https://doi.org/10.31763/businta.v6i1.593

Shaikh S, Daudpotta SM, and Imran AS (2021). Bloom's learning outcomes' automatic classification using LSTM and pretrained word embeddings. IEEE Access, 9: 117887-117909. https://doi.org/10.1109/ACCESS.2021.3106443

Supraja S, Hartman K, Tatinati S, and Khong AW (2017). Toward the automatic labeling of course questions for ensuring their alignment with learning outcomes. In the Proceedings of the 10th International Conference on Educational Data Mining, Wuhan, China: 56-63.

Ujkani B, Minkovska D, and Stoyanova L (2021). Using natural language processing for quality assurance purposes in higher education. In the 2021 IV International Conference on High Technology for Sustainable Development (HiTech), IEEE, Sofia, Bulgaria: 1-4. https://doi.org/10.1109/HiTech53072.2021.9614206

Valenti S, Neri F, and Cucchiarelli A (2003). An overview of current research on automated essay grading. Journal of Information Technology Education: Research, 2(1): 319-330. https://doi.org/10.28945/331

Zaki N, Turaev S, Shuaib K, Krishnan A, and Mohamed E (2023). Automating the mapping of course learning outcomes to program learning outcomes using natural language processing for accurate educational program evaluation. Education and Information Technologies, 28(12): 16723-16742. https://doi.org/10.1007/s10639-023-11877-4