Leveraging Machine Learning in Student Peer Review: A Systematic Literature Review

Theresia Devi Indriasari[1], Yohanes Sigit Purnomo W.P.[2]
Program Studi Informatika, Fakultas Teknologi Industri, Universitas Atma Jaya Yogyakarta
Jalan Babarsari 43, Yogyakarta 55281, Indonesia
Email: [1]devi.indriasari@uajy.ac.id , [2]sigit.purnomo@uajy.ac.id

Abstrak. *Penelitian ini mengkaji bagaimana teknik pembelajaran mesin diintegrasikan dalam proses student peer review, dengan berfokus pada tantangan yang mendorong penerapannya serta metode yang digunakan untuk mengatasinya. Dengan menggunakan kerangka Systematic Literature Review dari Kitchenham, sebanyak 328 artikel diseleksi menjadi 25 studi empiris yang membahas penerapan pembelajaran mesin dalam student peer review. Hasil penelitian menunjukkan bahwa pembelajaran mesin terutama digunakan untuk mengelola volume ulasan yang besar, mendukung penilaian otomatis, dan meningkatkan kualitas umpan balik. Teknik yang umum digunakan meliputi klasifikasi, prediksi, pemeringkatan, dan pengelompokan yang berkontribusi terhadap peningkatan akurasi, efisiensi, dan objektivitas dalam proses peer review. Penelitian ini menyajikan sintesis yang sistematis mengenai penerapan pembelajaran mesin dalam student peer review serta menyoroti potensinya dalam meningkatkan akurasi penilaian, mendukung capaian pembelajaran, dan menjadi dasar bagi penelitian serta implementasi yang lebih luas dalam konteks pendidikan.*
Kata kunci: *Pembelajaran Mesin, Penilaian Sejawat, Student Peer Review, Tinjauan Pustaka Sistematis, Umpan Balik Otomatis*

Abstract. *Our study examines how machine learning techniques are integrated into student peer review processes, focusing on the challenges that motivate their adoption and the methods used to address them. Using Kitchenham's systematic literature review framework, 328 articles were screened, and 25 empirical studies on machine learning applications in student peer review were selected. The findings show that machine learning is mainly used to manage large volumes of reviews, support automated grading, and improve feedback quality. Common techniques include classification, prediction, ranking, and clustering, which help improve the fairness, efficiency, and objectivity of peer review. This study provides a rigorous synthesis of machine learning adoption in student peer review and highlights its potential to enhance assessment accuracy, support learning outcomes, and guide future research and broader implementation in educational contexts.*
Keywords: *Automated Feedback, Machine Learning, Peer Assessment, Student Peer Review, Systematic Literature Review*

1. Introduction

Peer review is a collaborative learning strategy in which students evaluate the quality, value, or performance of their peers' work, typically among individuals of equal status [1]. It has been widely adopted in higher education to support learning, reflection, and assessment across disciplines [2]. Peer review benefits both instructors and students. For instructors, it can increase the amount of feedback students receive while reducing the burden of providing individual feedback, especially in large classes [3]. For students, peer review promotes active learning, communication, critical thinking, and the ability to give and receive constructive feedback[4]. It also encourages students to reflect on the quality of their own work and to make improvements based on peers' feedback [5].

Despite these benefits, student peer review still faces several challenges. Student reviewers may provide inconsistent scores, superficial comments, biased evaluations, or feedback that does not align with the assessed work [6]. These problems are often caused by differences in

students' knowledge, review skills, and seriousness about completing the assessment task. In large classes, manually checking the quality and consistency of peer feedback also becomes impractical for instructors [7], [8]. These issues highlight the need for computational methods that can support the management and improvement of peer review.

To address these challenges, researchers have increasingly applied machine learning to student peer review. Existing studies have used machine learning to classify peer feedback, detect inconsistencies between scores and comments, predict essay rankings, filter biased scores, evaluate reviewer reliability, and provide automated feedback to students [9], [10]. These applications suggest that machine learning can support peer review by improving efficiency, consistency, and scalability.

Given the growing use of machine learning in student peer review, a systematic literature review is needed to synthesize current findings, identify research gaps, and clarify the practical relevance of this area for different stakeholders. Therefore, this study conducts a systematic literature review guided by Kitchenham's SLR framework to examine the motivations, methods, and outcomes of machine learning applications in student peer review, as well as areas requiring further investigation. Although machine learning has been widely studied in broader educational contexts, reviews focusing specifically on student peer review remain limited. This review identifies empirical studies that apply machine learning to improve peer review processes and synthesizes their motivations, methods, and findings. The results are expected to benefit educators seeking to improve the quality and consistency of peer assessment, instructional designers developing effective peer review activities, educational technology developers designing scalable and fair review systems, and researchers advancing evidence-based approaches to technology-enhanced assessment. Overall, the findings provide insights into how machine learning can address key challenges in student peer review, particularly in improving feedback quality, fairness, consistency, efficiency, and scalability.

2. Related Work

Systematic literature reviews (SLRs) are widely recognized as rigorous and replicable methods for identifying, evaluating, and synthesizing relevant studies on a focused research question. Compared with ad hoc narrative reviews, SLRs provide greater transparency, methodological control, and reliability in summarizing existing evidence [11], [12], [13]. They typically involve protocol-driven searches across multiple sources, systematic integration of findings, and critical appraisal of the scope, nature, and quality of the evidence. Through this process, SLRs support broader theoretical understanding and practical implications [14], [15]. In evidence-based software engineering, Kitchenham's work has been particularly influential in positioning SLRs as original research methods that can replace informal reviews by producing fair, rigorous, and auditable summaries of evidence [16].

Kitchenham's SLR protocol is commonly organized into three phases: planning, conducting, and reporting the review [16]. The planning phase requires researchers to define clear research questions and prepare a review protocol that specifies the search strategy, inclusion and exclusion criteria, quality assessment procedures, and data extraction and synthesis methods. During the conducting phase, researchers search relevant digital libraries, screen studies using predefined criteria, assess study quality, and extract data for narrative or quantitative synthesis. Although this process can be time-consuming and methodologically challenging, especially when studies use diverse empirical approaches, it aims to produce a comprehensive and unbiased synthesis of evidence. The reporting phase emphasizes transparent documentation of each step so that the review process, potential bias, and credibility of the findings can be assessed and used to inform future research and practice.

3. Methods

Our systematic literature review was conducted in accordance with the guidelines outlined by [16]. This section provides a thorough explanation of every stage of the research

project, including the research questions, the search strategy, the chosen databases, and the techniques used for data extraction and analysis.

3.1. Research Questions

The objective of this investigation is to identify how ML is applied to address specific issues or challenges encountered in implementing peer review within educational settings. In this investigation, we address the following research questions: (1) RQ1: What specific challenges in peer review have motivated the adoption of machine learning, and what ML techniques have been employed to address these issues? (2) RQ2: How have different machine learning approaches been developed and tested to improve student peer review processes?

3.2. Search Process

We constructed a search query consisting of two main terms, namely ”peer review” and ”machine learning,” aligning with the central focus of our study. Each primary term was expanded by incorporating several synonymous expressions. Variants for ”peer review” included ”peer feedback,” ”peer assessment,” and ”peer grading,” while equivalents for ”machine learning” encompassed ”artificial intelligence.” In the subsequent step of the process, these search terms were interconnected using logical operators, employing ”AND” for the three primary components of the search string and ”OR” for synonymous phrases and keywords. The final search string (”peer review” OR ”peer feedback” OR ”peer assessment” OR ”peer grading”) AND (”machine learning” OR ”artificial intelligence”) was then executed across the ACM Digital Library (ACM DL), IEEE Xplore, and Science Direct databases. The search process was conducted in May 2023. Minor adjustments to the search strategy were implemented for each database due to its unique characteristics. The conditions specified for each respective database are as follows: (1) ACM Digital Library: search in the abstract, (2) IEEE Xplore: search in all, (3) Science Direct: search in title, abstract, or author-specified keywords.

Table 1. Inclusion and exclusion criteria for the selection process

Inclusion Criteria	Exclusion Criteria
Singular titles were incorporated	Duplicated titles
Articles subjected to peer review	Articles not subjected to peer review
Composed in the English language.	Written in languages other than English.
Easily obtainable	Inaccessible
Studies exploring peer review in educational contexts (such as integration into the teaching process)	Studies not addressing peer review in educational settings
Empirical studies	Non-empirical studies
Utilization of ML or AI in student peer review activities (such as utilizing peer review data to develop models and implementing ML for clustering student feedback)	Do not incorporate ML or AI in student peer review activities
Explicit mention of the utilized ML or AI technique	No explicit mention of the utilized ML or AI technique

3.3. Selection Process

The primary studies were selected through a three-stage selection process. Initially, data from all 328 studies were compiled into a single spreadsheet. The first stage of selection involved eliminating duplicate titles. Afterward, the titles, keywords, and abstracts of the remaining papers were examined, focusing exclusively on English articles that had undergone the peer review process for scientific publication. As our focus is on examining the use of ML in student peer review, we specifically included articles related to peer review practices within educational settings. Articles addressing peer review in non-educational contexts, such as the software industry, healthcare, academic research, and scientific publishing, were excluded to ensure the relevance and applicability of findings to educational environments. We intentionally excluded literature reviews, focusing instead on empirical studies to explore the methods and findings

related to the use of ML in student peer review. However, articles with insufficient information in their keywords, abstracts, or titles were still included in the next stage for further evaluation.

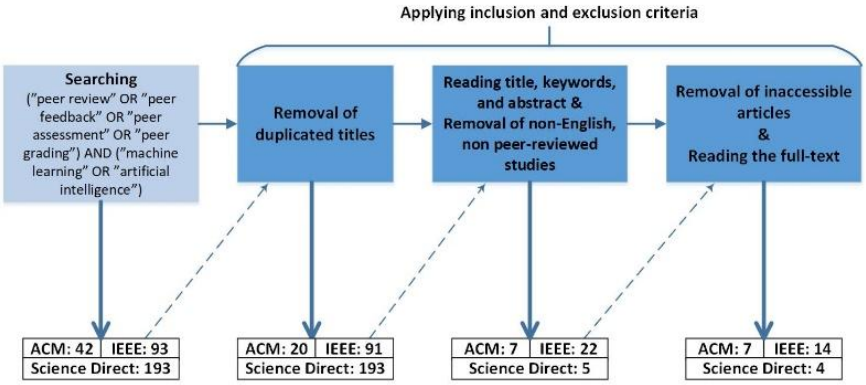

Figure 1. The selection process begins with the search and involves three subsequent steps.

In the third step of the selection process, a thorough analysis was conducted of the entire text of the remaining articles, excluding items that were not accessible. More specific criteria were applied, focusing on studies directly relevant to the research questions and presenting empirical data that implemented ML in student peer review activities. These criteria demanded sufficient detail regarding the ML technique employed. Table 1 provides a comprehensive overview of the inclusion and exclusion criteria at all stages. In addition, Figure 1 illustrates the search stage, which initially yields 328 studies. The first and second selection stages narrow this down to 34 studies, and the final selection stage further refines the selection to 25 articles for subsequent analysis and synthesis. The list of 25 primary studies is available in the first column of Table 2. The primary studies are identified in alphabetical order by article title.

3.4. Quality Assessment

Following Kitchenham's guidelines, we conducted a study quality assessment to ensure that the selected primary studies provided a sufficient basis for reliable synthesis. The assessment emphasized four main aspects: (1) clarity of research objectives and context, (2) transparency and sufficiency of methodological reporting to allow critical appraisal, even if not full replication, (3) appropriateness of the analysis techniques used, and (4) relevance of the reported outcomes to peer review in educational settings. Although several studies did not report complete details of their datasets or model development steps, they still met the minimum quality requirements when their aims were explicit, their methodological approach was understandable, and their findings were clearly linked to peer review challenges. All 25 studies satisfied these thresholds and were therefore retained in the synthesis. This approach ensures that the evidence base remains credible while acknowledging variations in reporting depth across individual studies.

3.5. Data Extraction and Data Synthesis

Relevant information related to our research questions was collected from primary studies using data extraction forms. An electronic form tailored specifically for data extraction was created and employed to capture the specific data outlined below: (1) Problem in Peer Review: Identifying the motivations behind the proposed ML solutions to address challenges in the peer review process. (2) ML Primary Tasks: Stating the primary tasks that the ML algorithms are designed to perform. (3) ML Algorithm: Identifying the specific ML algorithms employed in the studies. (4) Dataset: Identifying the datasets used in the studies. (5) ML Development: Identifying the cross-validation (CV) approaches, the evaluation metrics of top-performing models, and the determination of model implementation status.

After completing the data extraction process, we synthesized the data by identifying recurring themes to categorize the papers in addressing the research questions. For example, in addressing RQ1, we initially extracted the motivation for applying ML, the main ML tasks, and

the algorithm used from the abstract, introduction, or method sections of each paper. We applied thematic analysis [17] to categorize the motivations driving the development of ML. We used terminology from ML literature to classify the main ML tasks. For RQ2, we analyzed the methods and experimental setups sections and the results of each article.

4. Results

4.1. RQ1 - Challenges in Peer Review for the Adoption of Machine Learning Techniques

RQ1 focuses on identifying the key challenges that have prompted the adoption of ML in student peer review processes, as well as the specific ML tasks used to perform them. Table 2 summarizes the results. We synthesize the challenges related to ML development and the main tasks of ML used in student peer review across primary studies. We included all the tasks mentioned in the article. Therefore, a primary study may have multiple tasks, as shown in Table 2.

We categorize the core problems or motivations that have driven the integration of ML into student peer review processes. The challenges are grouped into specific categories: (1) Automating the grading of large volumes of student work: this category highlights the need for automatic grading systems to efficiently handle large quantities of student submissions. (2) Exploring a large volume of peer reviews: this challenge refers to the difficulty of managing and analyzing a significant number of peer review comments or feedback submissions. In large classes or settings with extensive peer interaction, the sheer volume of data can overwhelm instructors, making it hard to derive meaningful insights. (3) The quality of feedback: this challenge addresses concerns regarding the variability, reliability, and effectiveness of the feedback provided by students during peer review. Feedback can be in the form of comments, scores, or both, and it plays a crucial role in guiding student improvement. (4) Others: this category includes other unique or less common challenges that do not fall under the primary categories listed above.

We categorized the primary tasks of ML algorithms used into several categories [18], [19]: (1) Regression (Re): prediction, often referred to as regression in the context of machine learning, involves estimating a continuous outcome variable based on one or more input variables. The goal is to model the relationship between the input (feature) values and the continuous output. (2) Classification (Cla): classification involves assigning input data to one of several predefined categories or groups. It is a supervised learning technique in which a model is trained on labeled data to classify new inputs into specific categories. (3) Ranking (Ra): ranking refers to the process of arranging items based on specific criteria, such as relevance, importance, or quality. (4) Semantic Analysis (SA): semantic analysis refers to the process of extracting meaning from text data. Often, it is used to help interpret the content and context of the text. (5) Clustering (Clu): clustering is an unsupervised learning task that involves grouping data points into clusters based on their similarities. It is used to identify natural groupings within the data without the need for labeled examples. (6) Outlier Detection (OD): outlier detection, also known as anomaly detection, involves identifying data points that significantly differ from the majority of the dataset. It is essential for detecting irregularities or errors. (7) Dimensionality Reduction (DR): dimensionality reduction involves reducing the number of variables or features in a dataset while retaining the essential information. This task is useful for simplifying models, removing noise, and visualizing high-dimensional data.

Table 2. The challenges that have motivated the adoption of ML and the methods used to implement it.

Primary Studies	Motivation	ML Task						
		Re	Cla	Ra	SA	Clu	OD	DR
ID1 [7]	Exploring a large volume of peer reviews		x					
ID2 [20]	Exploring a large volume of peer reviews		x					
ID3 [8]	Exploring a large volume of peer reviews		x					

ID4 [21]	Automating the Grading of Large Volumes of Student Work		x					
ID5 [22]	The quality of feedback	x						
ID6 [23]	others			x				
ID7 [24]	The quality of feedback		x					
ID8 [25]	The quality of feedback		x					
ID9 [26]	The quality of feedback	x						
ID10 [27]	The quality of feedback				x			
ID11 [28]	The quality of feedback	x	x				x	
ID12 [29]	The quality of feedback	x						
ID13 [30]	The quality of feedback			x				
ID14 [31]	Exploring a large volume of peer reviews	x						
ID15 [32]	Exploring a large volume of peer reviews					x		
ID16 [33]	The quality of feedback		x					
ID17 [34]	Exploring a large volume of peer reviews					x		
ID18 [35]	The quality of feedback			x				
ID19 [36]	others		x					
ID20 [37]	The quality of feedback	x		x				
ID21 [38]	The quality of feedback		x					
ID22 [39]	Exploring a large volume of peer reviews		x					
ID23 [40]	Automating the Grading of Large Volumes of Student Work	x						
ID24 [41]	Automating the Grading of Large Volumes of Student Work	x						
ID25 [42]	Exploring a large volume of peer reviews		x		x			x

Table 2 highlights that the most commonly reported issues in the student peer review process (noted in 12 out of 25 articles) were associated with feedback quality, including the scores, the comments, or both. Throughout the studies examined, various aspects of feedback quality were identified as critical areas needing improvement, with machine learning (ML) approaches developed to address these concerns. Several studies have concentrated on enhancing the fairness and accuracy of student reviews by targeting inconsistencies, biases, and the overall reliability of scoring. For instance, ID5 focuses on identifying mismatches between numerical scores and written comments to ensure that feedback accurately represents the assessment of student work. Likewise, studies like ID9, ID11, ID12, and ID13 work to filter out biased or random scores that may arise from students who may be inexperienced, less committed, or have varying levels of knowledge. These efforts are aimed at improving the fairness and precision of student grading. In addition, ID18, ID20, and ID21 address the challenge of ensuring consistent and fair peer reviews by exploring methods for aggregating imperfect grades to produce a more balanced final assessment.

Several studies have also focused on the quality of comments provided in peer reviews. For example, ID7 and ID8 examine methods to give students immediate feedback on the quality of their reviews, with the aim of reducing poorly written reviews while encouraging more constructive input. ID16, meanwhile, evaluates the usefulness of feedback by recognizing the variation in quality and emphasizing comments that offer meaningful value to the recipient. Additionally, ID10 addresses the issue of harmful feedback, such as disrespectful or bullying remarks, aiming to protect students and prevent disengagement from the learning process.

From a teacher's perspective, one of the key challenges in peer review is managing and extracting valuable insights from large amounts of feedback. This difficulty has motivated the

adoption of machine learning techniques in eight of the 25 studies reviewed. For instance, ID1 and ID3 highlight the challenges instructors face when manually sifting through extensive peer review comments, highlighting the need for tools that can effectively extract meaningful insights. To address this, these studies utilize natural language processing (NLP) methods, like sentiment analysis, to classify and evaluate student feedback, allowing teachers to quickly access and apply the insights provided. In the context of second language (L2) writing, ID2 examines the types and characteristics of peer feedback and how they influence revisions. In ID2, machine learning enhances the analysis of web-based peer review comments. Similarly, ID14 explores how students' different personality profiles can shape their subjective evaluations of peer reviews. ID15 and ID17 focus on identifying patterns within student feedback to provide tailored, meaningful feedback and improve instructional strategies for students. ID22 applies sentiment analysis to forum messages, providing a glimpse into student emotions and engagement in MOOCs. Finally, ID25 analyzes linguistic patterns in social learning platforms to classify the dialogic functions of student interactions.

Meanwhile, three out of 25 studies highlight the problem of automatically grading large volumes of student work in peer review activities. ID4 addresses this by developing a platform that uses Support Vector Machines (SVM) to automatically check whether students' arguments meet the expected learning objectives, thus reducing the manual grading burden on teachers. ID23 introduces the K-OpenAnswer system, which semi-automates grading by combining peer reviews with selective teacher input, making it feasible to assess open-ended assignments in large courses like MOOCs. Similarly, ID24 implements ML for automated essay scoring, allowing instructors to maintain grading quality despite having limited resources.

Lastly, two studies highlight other peer review challenges that motivated the adoption of machine learning (ML). ID6 tackles the limitations of traditional ranked peer grading algorithms, developing the BayesRank model to enhance the accuracy and efficiency of ranking student work. ID19 addresses the issue of poorly understood rubrics by proposing an automated rubric analysis approach based on NLP. This approach evaluates whether rubric items are likely to encourage reviewers to provide high-quality feedback, thus improving the clarity and effectiveness of the assessment criteria.

The primary studies have adopted various ML tasks and algorithms to address the above peer review challenges. Notably, several studies implement more than one ML task, each applying different algorithms to tackle the multifaceted problems in the peer review process. Classification is the most commonly applied ML task, with numerous studies employing various algorithms to categorize textual data and student submissions. ID1 uses RoBERTa to classify peer review comments into categories such as Problem/Solution, Praise, or Verification/Summary, enhancing the clarity and utility of feedback. ID2 applies Logistic Regression to predict whether peer feedback will prompt a revision in the student's draft, while ID3 leverages Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), Logistic Regression, and Random Forest (RF) to classify the sentiment of feedback as positive or negative. ID4 uses SVM to classify whether students addressed relevant learning issues on discussion boards, ensuring the alignment of peer assessments with course objectives. ID7 and ID8 concentrate on automating the classification of peer review comments. ID7 utilizes Convolutional Neural Networks (CNN), while ID8 applies BART for this task. ID11 combines several algorithms, including Naive Bayes, Long Short-Term Memory (LSTM), and CNN, to classify student submissions and detect outlier grades. ID16 utilizes a comprehensive set of algorithms, including Logistic Regression, RF with various boosting techniques, and DistilBERT, to classify the helpfulness of peer feedback. ID19 utilizes several algorithms, including SVM, RF, Gradient Boosting, BiLSTM, LSTM, CNN+LSTM, and BERT, to classify review comments and detect rubric items that enable high-quality feedback. ID21 applies a generic text classifier to categorize student short-answer submissions into predicted grades. ID22 uses a suite of algorithms, including Logistic Regression, SVM, Decision Trees, Naïve Bayes, and Lexicon-Based approaches like SentiWordNet to analyze the sentiment and polarity of forum messages, thus providing insights into student emotions and engagement. Finally, ID25 implements Latent Dirichlet Analysis (LDA) and SVM to classify student and

teacher messages into different dialogic functions, enhancing the understanding of communication patterns in peer review.

Prediction or regression tasks play a role in estimating continuous variables, such as grades and behavioral intentions. ID5 applies a variety of models, including k-Nearest Neighbors (kNN), SVM, Decision Trees, RF, and Neural Networks, to predict numerical scores from textual feedback. ID9 integrates prediction models like Hammer's and MACE to estimate accurate grades from peer review. ID11 uses Bayesian models to predict the actual grades of student submissions by analyzing peer grades. ID12 employs algorithms like Linear Regression, SVM, and RF to predict the seriousness of graders based on their behaviors. ID14 uses regression techniques such as Linear Regression, Lasso, Ridge, and Decision Trees to predict behavioral outcomes based on personality traits. ID20 combines supervised and unsupervised methods to predict grading inaccuracies and solve ranking problems. ID23 utilizes kNN to predict student performance and infer grades based on peer evaluations. ID24 applies RF to predict assessment scores for student writing in large classes.

Ranking tasks focus on ordering student submissions according to their quality or relevance. ID6 introduces the BayesRank model to rank student work based on peer reviews and to address the limitations of previous ranked peer grading systems. ID13 and ID18 employ SGD and Matrix Factorization to optimize the ranking of assignments and essays and to enhance the accuracy of peer review outcomes. ID20 also addresses ranking through Bayesian models, predicting true grades, and ranking submissions accordingly.

Clustering tasks are applied to group similar data points, such as feedback comments or student submissions. ID15 uses k-Means and Convex Optimization to capture subjective similarity measurements, thus facilitating the identification of common themes in peer feedback. ID17 applies a Gaussian Mixture Model (GMM) to cluster student responses that can provide instructors with actionable insights into student performance.

Semantic Analysis tasks are utilized to extract meaning from textual data, thus helping to understand the content and context of peer feedback. ID10 uses Latent Semantic Analysis (LSA) to distinguish between relevant and irrelevant comments. ID25 applies LDA for topic modeling to uncover latent topics in the dataset.

Outlier Detection is a task for identifying anomalies in peer grading. ID11 uses kNN to detect outlier grades, ensuring that extreme or biased scores do not skew the overall assessment. Lastly, Dimensionality Reduction is employed in ID25 to simplify complex datasets through Principal Component Analysis (PCA), making it easier to analyze and interpret the multidimensional data found in peer review comments.

4.2. RQ2 - Development and Evaluation of Machine Learning Approaches to Enhance Student Peer Review Processes

The RQ2 aims to explore the development and assessment of ML approaches to enhance student peer review processes. To explore the ML approaches in each study, first, we identify the main task of ML, followed by an analysis of the datasets used, and finally, how the ML models were developed. The main task of ML contains the peer-review-related tasks that ML has solved. The dataset includes details such as 1) the dataset size; 2) the data source, which could be students (S), teachers (T), experts (E), or teaching assistants (TA); 3) the type of peer review data, categorized as student review (R) or student work (W); 4) the type of data, that is classified into text and scores or grades (S/G)); and lastly 5) the language (LANG) used, including English (EN), Chinese (ZH-CN), Spanish (ES), and Chilean Spanish (ES-CL).

The ML development section focuses on the cross-validation (CV) approaches used during the development, the evaluation metrics of the top-performing models, and whether the models were implemented in practice. Cross-validation, especially for smaller datasets, is a critical step that involves repeatedly training and testing the models on various subsets or random divisions of the dataset. For each column in the table, if information is unclear or not provided by the primary studies, it is noted as Not Available (NA) in the table. We then grouped the results based on the ML task from Table 2. We present them in four tables, including Table 3 for the

regression tasks, Table 4 for the classification tasks, Table 5 for the ranking tasks, and Table 6 for other tasks.

Table 3. The machine learning approach for the regression tasks.

ID	TASK	DATASET					MACHINE LEARNING DEVELOPMENT		
		SIZE	SOURCE	PR DATA	TYPE OF DATA	LANG	CV	Best Model Performance	Model Impl.
ID5	to predict the numerical score from the textual feedback	956	S	W	text	EN	10-fold	MAE of 0.22 (LSTM with 2-g)	No
		969	S	W	text	EN	10-fold	MAE of 0.23 (LSTM with 2-g)	No
ID9	to predict the true grades of student submissions	1714	S	R	S/G	NA	NA	RMSE of 1.290994 (Proposed algorithm)	No
ID11	to infer the true scores	1260	S	W	text	ZH-CN	5-fold	RMSE of 11.32 (PG4)	No
ID12	to predict the grading seriousness	1279	S	R	S/G	NA	LOSO	RMSE of 1.35 (Random Forest)	No
		1297	S	R	S/G	NA	LOSO	RMSE of 1,65 (GDBT)	No
		1526	S	R	S/G	NA	LOSO	RMSE of 1,57 (GDBT)	No
	to predict the true scores	1279	S	R	S/G	NA	NA	RMSE of 1.22 (BPG6)	No
		1297	S	R	S/G	NA	NA	RMSE of 1.99 (BPG7)	No
		1526	S	R	S/G	NA	NA	RMSE of 1.36 (BPG6)	No
ID14	to predict various behavioral intention outcomes based on personality traits and the modality of peer assessment	1338	S, E	R	NA	NA	10-fold	IU: RMSE of 7,1 (RF), U: RMSE of 5.9 (RF), EU: RMSE of 5.7 (RF), C: RMSE of 5,7 (RF, XGB)	No
ID20	to predict the true grades of student submissions and rank them accordingly	NA	S, TA	R	S/G	NA	NA	NA	No
	to predict inaccuracies	NA	S, TA	R	S/G	NA	NA	NA	No
ID23	to predict student performance	4000	S	R	S/G	NA	NA	NA	Yes
ID24	to predict assessment scores	196	S	W	text	EN	10-fold	86.7% accuracy (RF)	Yes

Table 3 provides some details of machine learning models for regression tasks within the peer review context. The datasets used for these tasks vary, ranging from smaller sets with around

196 records (ID24) to larger datasets with over 12,000 records (ID20). The tasks generally involve predicting numerical outcomes, such as the true grades of student submissions (ID5, ID9, ID11, ID12, ID20, ID23, and ID24), grading seriousness (ID12), and behavioral intention outcomes based on various input features like textual feedback, grading behaviors, and personality traits (ID14). The performance of these machine learning approaches is assessed using metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), which provide a quantitative measure of the prediction accuracy. For instance, LSTM models achieved a low MAE of 0.22 and 0.23 when predicting numerical scores from textual feedback, indicating their effectiveness in this context (ID5).

On the other hand, the proposed algorithm for predicting true grades of student submissions yielded an RMSE of approximately 1.29 (ID9), suggesting room for improvement in this application. In addition, ID20 shows that the dataset size is NA because the study does not mention the number of grades used in the machine learning development; the performance of the developed ML models is also filled with NA because this study uses a boxplot chart to show Kendall-Tau metrics, but the value of the best model is not clear. During the model development, some studies use cross-validation, such as k-fold (ID5, ID11, ID14, ID24) and LOSO (Leave-One-Subject-Out) (ID12). Most of the ML models in regression tasks have not been implemented yet, and only two studies have implemented the ML model (IDs 23 and 24).

Table 4. The machine learning approach for the classification tasks.

ID	TASK	DATASET					MACHINE LEARNING DEVELOPMENT		
		SIZE	SOURCE	PR DATA	TYPE OF DATA	LANG	CV	Best Model Performance	Model Impl.
ID1	to classify comments	6431	S	R	text	EN	NA	MCC score of 0.687 (RoBERTa)	Yes
		4372	S	R	text	EN	NA	MCC score of 0.61 (RoBERTa) for Incorrect Content	Yes
ID2	to classify whether feedback will result in a revision or not	885	S	R	text	EN	10-fold	AUC of 0.617 ± 0.055 (Logistic Regression)	No
	to predict which attributes are most relevant for the revision	885	S	R	text	EN	10-fold	AUC of 0.641 ± 0.050 (Logistic Regression without kfold)	No
ID3	to classify the sentiment of feedback	3968	S	R	text	ES	NA	F1 score of 0.8785 (SVM with 1-grams+2-grams+TFIDF)	No
ID4	to classify whether students addressed the related learning issues	1258	S	W	text	ZH-CN	LOOCV	recognition rates (91%)	Yes
ID7	to give automated feedback by	NA	NA	NA	NA	NA	NA	NA	Yes

	using labeled data								
ID8	to classify peer review moves	692	S	R	text	EN	NA	weighted avg. f1-score of 0.50 (BART)	No
		538	S	R	text	EN	NA	weighted avg. f1-score of 0.22 (BART)	No
		538	S	R	text	EN	NA	weighted avg. f1-score of 0.68 (BART)	No
		538	S	R	text	EN	NA	weighted avg. f1-score of 0.77 (BART)	No
ID11	to predict the numerical score from the submission	1260	S	W	text	ZH-CN	5-fold	accuracy of 82.76 - 86.54 % (CNN)	No
ID16	to classify subjective opinions (helpfulness)	44470	S	R	text	EN	5-fold	F1 score of 0,73 (Text CNN, LSTM)	No
	to classify subjective opinions (helpfulness) with human knowledge	44470	S	R	text	EN	NA	F1 score of 0.753 (Augmented DistilBERT)	No
ID19	to classify review comments	40814	S	R	text	EN	NA	Detects Problem: 0.91 F1 score (BERT), Gives Suggestion: 0.91 F1 score (BERT), and Localization: 0.8 (BERT, CNN+LSTM, Bi-LSTM)	No
	to classify rubric items that enable peer reviewers to write quality reviews	40814	S	R	text	EN	NA	0.86 PRAUC score (BERT)	No
ID21	to classify student short-answer submissions	500	S, TA	R	text and S/G	EN	NA	NA	Yes
ID22	to classify the polarity (positive or negative)	500	S	R	text	EN	LOOCV	0.71 AUC (RF, Dictionaries)	No
		134	S	R	text	EN	LOOCV	0.85 AUC (NB), 0.78 (Dictionaries)	No
ID25	to classify messages	500	S	W	text	ES-CL	10-fold	92,6% F1 score all (SVM)	No

For the classification tasks, as shown in Table 4, most studies used ML to classify the students' feedback. The dataset sizes vary significantly across different tasks, ranging from smaller datasets of a few hundred samples (e.g., 500 to 885 records) to larger datasets with tens of thousands of records (e.g., 44,470 to 40,814 records). Most of these datasets consist of text data in English, though some tasks utilize data in other languages like Spanish (ID3), Chilean Spanish (ID25), and Chinese (ID4 and ID11). The development of the machine learning models often involves using cross-validation techniques such as 5-fold (ID11 and ID16) or 10-fold CV (ID2 and ID25), leaving one out (ID4 and ID22), though some tasks lack cross-validation information. Specific to ID16, transformer-based models, such as Augmented DistilBERT, did not use cross-validation because of the nature of the transformer-based model. The models achieve varying degrees of success, with performance metrics such as MCC score, AUC, F1-score, and accuracy indicating the efficacy of different approaches. ID19 uses the PRAUC score because of the imbalanced data. In ID21, the researcher uses the website www.etcml.com to develop a generic text classifier with the predicted grade as the output to reduce human effort and integrate it into a peer review that incorporates human and machine grading. Notably, just four studies have been successfully implemented and deployed (ID1, ID4, ID7, and ID21), suggesting their practical viability in enhancing peer review processes. For ID7, most columns are filled with NA because it is unclear how the dataset is used in ML model development, that is, a CNN. In this study, both CNN models achieved a label certainty threshold of 0.9, enabling them to infer over 50% of the labels that students initially had to compute.

Table 5. The machine learning approach for the ranking tasks.

ID	TASK	DATASET						MACHINE LEARNING DEVELOPMENT	
		SIZE	SOURCE	PR DATA	TYPE OF DATA	LANG	CV	Best Model Performance	Model Impl.
ID6	to rank student work based on peer assessments	2632	S	R	S/G	NA	NA	Kendall's-T of 0.2971 (BayesRank)	No
		3420	S	R	S/G	NA	NA	Kendall's-T of 0.4463 (BayesRank)	No
		2880	S	R	S/G	NA	NA	Kendall's-T of 0.4373 (BayesRank)	No
		2438	S	R	S/G	NA	NA	Kendall's-T of 0.4441 (BayesRank)	No
ID13	to rank the assignments according to the partial orders given by the graders	1326	S	R	S/G	NA	NA	AUC score of 0.830 (SGD)	No
ID18	to rank English essays	100	S	R	S/G	NA	NA	loss of 13.61% (SGD)	No
		200	S	R	S/G	NA	NA	loss of 15.29% (SGD)	No
		177	S	R	S/G	NA	NA	loss of 02.10% (SGD)	No
ID20	to predict the true grades of student submissions and rank them accordingly	NA	S, TA	R	S/G	NA	NA	NA	No

Table 5 shows machine learning approaches for ranking tasks in peer review. For instance, BayesRank is frequently used to rank student work based on peer assessments, achieving varying performance levels as measured by Kendall's Tau coefficients (ID6). These coefficients indicate the correlation between the predicted and actual rankings, with higher values representing better performance. Additionally, SGD is used for tasks involving ranking English essays and predicting true grades, with performance metrics expressed as AUC score (ID13) and loss percentages (ID18). The loss values provide insight into the model's ability to minimize errors during prediction, with lower loss percentages indicating better performance.

Table 6. The machine learning approach is used for the other tasks, such as semantic analysis, clustering, outlier detection, and dimensionality reduction.

ID	TASK	DATASET					MACHINE LEARNING DEVELOPMENT		
		SIZE	SOURCE	PR DATA	TYPE OF DATA	LANG	CV	Best Model Performance	Model Impl.
ID10	to evaluate peers' feedback	> 1000	S	R	text	EN	NA	NA	No
ID11	to detect outlier grades in the peer grading process	1260	S	W	text	ZH-CN	5-fold	RMSE of 20.35 - 22.09 (KNN)	No
ID15	to capture users' subjective similarity measurements	NA	NA	NA	NA	NA	NA	NA	Yes
ID17	to automatically provide some feedback to the instructors (topic modeling)	1326	S	R	S/G	NA	NA	BIC of -6245.4 and AIC of -7080.8 (GMM)	No
		175	S	W	text	EN			
		1065	S	R	S/G	NA	NA	BIC of -16569.9 and AIC of -18043.4 (GMM)	No
		111	S	W	text	EN			
		660	S	R	S/G	NA	NA	BIC of -13777.2 and AIC of -15299.0 (GMM)	No
		66	S	W	text	EN			
ID25	to obtain insight into the content of student and teacher messages to help us understand linguistic attitudes	692	S	W	text	ES-CL	NA	NA	No
		101	T	W	text	ES-CL			

Finally, Table 6 shows the machine learning approach used in peer review for semantic analysis, clustering, outlier detection, and dimensionality reduction. As demonstrated in Table 6, models like KNN is employed to solve the outlier detection task, with RMSE values indicating the effectiveness of the model in identifying outliers in peer grading (ID11). In ID15, most columns are marked as NA because the study primarily focuses on the use of a tool, MindMiner,

to capture domain experts' subjective similarity measurements through a combination of novel interaction techniques and machine learning algorithms. Additionally, for the clustering tasks, GMMs are utilized, with performance metrics such as BIC and AIC scores indicating the model's fit and complexity (ID17). ID25 uses LDA for topic modeling, and PCA is applied as an exploratory data analysis to discover features in the multidimensional data.

5. Discussion

Peer review is a well-established pedagogical tool that supports the development of critical thinking, communication skills, and self-assessment among students[43]. Rooted in constructivist learning theories, peer review encourages active engagement and collaboration, potentially deepening students' understanding of the material [44]. However, its effectiveness is closely tied to clear guidelines and appropriate training. Guidelines such as rubrics can help students understand the evaluation criteria, improving the consistency and quality of their feedback[45]. In the absence of these, peer reviews can be prone to inconsistencies and biases, which may reduce their educational benefit [46]. Maintaining objectivity in peer review is also crucial, and strategies like anonymity and involving multiple reviewers can help reduce bias [47]. However, despite these efforts, peer review can still have challenges, especially if the reviewers lack sufficient knowledge and skill. This can result in feedback that is less constructive than intended, potentially being vague, overly critical, or even inaccurate. From the perspective of the recipients of the feedback, such responses can be less useful or discouraging. When students receive unclear or harsh feedback, they may start to question their abilities, which can result in a loss of confidence. This loss of confidence can make them less engaged and hinder their overall learning outcomes [48].

Building on these points, the findings from RQ1 highlight several challenges that have led to the adoption of ML in peer review processes. Specifically, the results indicate significant issues in the student peer review process, particularly related to the quality and consistency of feedback provided. Machine learning techniques have been developed to help address these challenges by potentially enhancing the fairness, accuracy, and reliability of peer assessments. For example, various ML algorithms have been used to detect inconsistencies between review scores and comments, filter out biased or inconsistent assessments, and improve the aggregation of peer-assigned grades [22], [28], [29], [30], [35], [37], [38]. These approaches may help mitigate some of the drawbacks in traditional peer review processes by making feedback more consistent, objective, and reflective of actual performance. This objectivity and accuracy play an important role in offering students feedback that better reflects their abilities and areas for growth. As a result, the adoption of ML in peer review has the potential to foster more meaningful learning experiences, which could contribute to improved educational outcomes for students.

The role of ML in enhancing peer review is further demonstrated through its ability to improve the analysis of feedback. Classification algorithms, for instance, can categorize feedback into meaningful groups, helping educators focus on the most valuable input [7]. When teachers focus on the most valuable comments, they may be better positioned to guide students more effectively, offering targeted support and interventions as needed. Additionally, ML can provide real-time feedback to students who act as reviewers, which might enhance their engagement and improve the quality of their reviews [24]. By offering instant analysis and feedback, ML-supported peer review processes can help students identify areas for improvement more quickly, leading to a more responsive learning experience. Furthermore, categorized feedback can encourage students to reflect more thoroughly on their work. It helps them distinguish between different types of input, such as praise, suggestions, or critiques. This allows them to adapt their learning strategies more effectively, to understand the material better, and to improve their critical thinking. Over time, these improvements could lead to better student outcomes.

From an instructor’s perspective, incorporating ML into student peer review processes has meaningful implications. ML offers a new approach to student peer review that can analyze large amounts of data with high accuracy and reveal insights that might be missed during manual reviews. This could advance educational theories of assessment, feedback, and learning

outcomes, particularly by exploring how automated processes can effectively complement human judgment. On a practical level, ML can significantly reduce the workload for instructors by automating the analysis and classification of peer feedback, enabling them to concentrate on more strategic aspects of teaching and learning. For example, instructors can devote more time to curriculum development, personalized instruction, and addressing individual student needs [34], [35]**.** In addition, by using ML to automate tasks, specific peer review challenges are addressed, resulting in a more scalable and efficient system.

Regarding the results from RQ2, the performance of ML models in the primary studies shows a wide range of results depending on the models and tasks involved. For instance, BERT models demonstrated strong performance with an F1 score of 0.91 for detecting problems and suggestions in peer review comments [36]. In addition, LSTM models used to predict numerical scores from textual feedback showed relatively low MAE values of around 0.22-0.23 [22], indicating good predictive accuracy. However, in other tasks, such as predicting the true grades of student submissions, the RMSE reached 1.29 [26], suggesting that there is still room for improvement. These results suggest that certain models perform well for specific tasks while also highlighting the challenges some tasks face in reaching higher levels of accuracy. Researchers developing ML models for peer review should carefully consider the specific tasks involved and choose models that best fit the data and evaluation criteria. Different algorithms may perform better depending on the context. Therefore, selecting the right approach is key to achieving better outcomes.

There are a couple of areas that might benefit from further exploration. First, concerning the datasets used, particularly those involving student feedback in text format, most studies have focused on English-language datasets, with a few also utilizing Chinese and Spanish. Since these tasks often involve NLP, there could be value in developing and testing datasets in other languages [49], [50], especially in low-resource languages that are commonly used in educational contexts in Southeast Asia, such as Bahasa Indonesia. This might help to broaden the applicability of ML models in diverse educational settings. Second, many ML techniques have not yet been fully adopted or smoothly integrated into peer review systems. Future research could focus on developing ML models to build a comprehensive framework that integrates various ML tasks to enhance the student peer review process. These tasks could include predicting student scores or grades based on their work and peer feedback, classifying feedback by sentiment or other subjective opinions, and deriving insights from the feedback data. Implementing and integrating these models into peer review systems could be a helpful direction for further study.

We recognize potential threats that may influence the validity of our findings. First, regarding the selection process, it is possible that some relevant studies were not captured. To mitigate this risk, we designed comprehensive search queries that incorporated multiple synonyms of the main concepts and applied them across three major digital libraries (ACM Digital Library, IEEE Xplore, and ScienceDirect). In addition, search strategies were adapted to the specific indexing and search features of each database to maximize coverage and reduce the likelihood of omission. Second, in terms of data extraction and synthesis, the tasks were conducted collaboratively by both authors to ensure accuracy and consistency. Each author independently extracted key information from the primary studies before systematically comparing results. Any discrepancies were resolved through structured discussions until consensus was reached. This rigorous approach minimized individual bias and provided a balanced interpretation of the evidence. Furthermore, the application of thematic analysis and the alignment of classifications with established machine learning terminology strengthened the transparency and reliability of the synthesis, particularly in addressing RQ1 and RQ2.

6. Conclusion

This systematic literature review aimed to explore the application of ML in improving student peer review processes. After screening 328 articles, 25 primary studies were selected for analysis. The analysis focused on the challenges that have driven ML adoption, the methods used, and the outcomes achieved. The findings identified key challenges in peer review, such as

feedback quality and consistency, managing large volumes of reviews, and automating grading tasks. ML techniques, including classification, prediction, ranking, and clustering, appear to offer the potential for enhancing the fairness, accuracy, and efficiency of peer assessments, contributing to a more objective process that better reflects students' actual performance.

The findings offer significant implications for educators and researchers. For educators, integrating ML into peer review processes can reduce workload and enhance scalability and effectiveness. For researchers, this study highlights the need to further explore ML techniques in education, with a focus on refining algorithms and understanding their long-term impact on learning outcomes. Future research should investigate the integration of ML-driven peer review across diverse educational settings and explore innovative approaches to further enhance its effectiveness.

References

[1] J. Serrano-Aguilera *et al.*, "Using Peer Review for Student Performance Enhancement: Experiences in a Multidisciplinary Higher Education Setting," *Education Sciences*, 2021, doi: 10.3390/educsci11020071.

[2] N. Ardill, "Peer feedback in higher education: student perceptions of peer review and strategies for learning enhancement," *European Journal of Higher Education*, vol. 15, no. 4, pp. 696–721, Jan. 2025, doi: 10.1080/21568235.2025.2457466.

[3] N. T. Kerman, S. K. Banihashem, M. Karami, E. Er, S. van Ginkel, and O. Noroozi, "Online peer feedback in higher education: A synthesis of the literature," *Education and Information Technologies*, vol. 29, no. 1, pp. 763–813, Jan. 2024, doi: 10.1007/s10639-023-12273-8.

[4] B. Ortega-Ruipérez and J. M. Correa-Gorospe, "Peer assessment to promote self-regulated learning with technology in higher education: systematic review for improving course design," *Frontiers in Education*, 2024, doi: 10.3389/feduc.2024.1376505.

[5] A. Annasekaran, V. Rajasekar, R. M. P, and V. Kalaivani, "Peer-Reviewed Reflective Writing by Phase II Medical Students: A Mixed-Method Study," *Cureus*, vol. 17, 2025, doi: 10.7759/cureus.90679.

[6] A. Darvishi, H. Khosravi, A. Rahimi, S. Sadiq, and D. Gašević, "Assessing the Quality of Student-Generated Content at Scale: A Comparative Analysis of Peer-Review Models," *IEEE Transactions on Learning Technologies*, vol. 16, pp. 106–120, 2023, doi: 10.1109/tlt.2022.3229022.

[7] S. Hutt *et al.*, *Feedback on Feedback: Comparing Classic Natural Language Processing and Generative AI to Evaluate Peer Feedback*. Proceedings of the 14th Learning Analytics and Knowledge Conference, 2024. doi: 10.1145/3636555.3636850.

[8] K. J. Topping, E. F. Gehringer, H. Khosravi, S. Gudipati, K. Jadhav, and S. Susarla, "Enhancing peer assessment with artificial intelligence," *International Journal of Educational Technology in Higher Education*, vol. 22, 2025, doi: 10.1186/s41239-024-00501-1.

[9] P. C. Sauer and S. Seuring, "How to Conduct Systematic Literature Reviews in Management Research: A Guide in 6 Steps and 14 Decisions," *Review of Managerial Science*, vol. 17, pp. 1899–1933, 2023, doi: 10.1007/s11846-023-00668-3.

[10] J. Paul, P. Khatri, and H. K. Duggal, "Frameworks for Developing Impactful Systematic Literature Reviews and Theory Building: What, Why and How?," *Journal of Decision Systems*, vol. 33, no. 4, pp. 537–550, 2024, doi: 10.1080/12460125.2023.2197700.

[11] M. Azarian, H. Yu, A. Shiferaw, and T. Stevik, "Do We Perform Systematic Literature Review Right? A Scientific Mapping and Methodological Assessment," *Logistics*, vol. 7, no. 4, p. 89, 2023, doi: 10.3390/logistics7040089.

[12] G. Marzi, M. Balzano, A. Caputo, and M. M. Pellegrini, "Guidelines for Bibliometric-Systematic Literature Reviews: 10 Steps to Combine Analysis, Synthesis and Theory

Development," *International Journal of Management Reviews*, vol. 27, no. 1, pp. 81–103, 2025, doi: 10.1111/ijmr.12381.

[13] W. Bandara and R. Syed, "The Role of a Protocol in a Systematic Literature Review," *Journal of Decision Systems*, vol. 33, no. 4, pp. 583–600, 2024, doi: 10.1080/12460125.2023.2217567.

[14] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004, [Online]. Available: http://www.it.hiof.no/~haraldh/misc/2016-08-22-smat/Kitchenham-Systematic-Review-2004.pdf

[15] V. Clarke and V. Braun, "Thematic analysis," *The Journal of Positive Psychology*, vol. 12, no. 3, pp. 297–298, 2016, doi: 10.1080/17439760.2016.1262613.

[16] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 3, p. 160, 2021, doi: 10.1007/s42979-021-00592-x.

[17] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, Jul. 2009, doi: 10.1145/1541880.1541882.

[18] A. Dood, K. Das, Z. Qian, S. Finkenstaedt-Quinn, A. Gere, and G. Shultz, "A Dashboard to Provide Instructors with Automated Feedback on Students' Peer Review Comments," in *LAK23: 13th International Learning Analytics and Knowledge Conference*, in LAK2023. New York, NY, USA: Association for Computing Machinery, 2023, pp. 619–625. doi: 10.1145/3576050.3576087.

[19] D. A. J. Leijen, "A Novel Approach to Examine the Impact of Web-based Peer Review on the Revisions of L2 Writers," *Computers and Composition*, vol. 43, pp. 35–54, 2017, doi: 10.1016/j.compcom.2016.11.005.

[20] M. P. Ortega, L. B. Mendoza, J. M. Hormaza, and S. V. Soto, "Accuracy' Measures of Sentiment Analysis Algorithms for Spanish Corpus generated in Peer Assessment," in *Proceedings of the 6th International Conference on Engineering & MIS 2020*, in ICEMIS'20. New York, NY, USA: Association for Computing Machinery, 2020. doi: 10.1145/3410352.3410838.

[21] C.-J. Huang, Y.-W. Wang, S.-C. Chang, S.-Y. Lin, J.-H. Tseng, and J.-J. Jian, "Applications of data mining to an online argumentation based learning assistance platform," in *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, 2012, pp. 807–811. doi: 10.1109/SCIS-ISIS.2012.6505083.

[22] J. R. Rico-Juan, A.-J. Gallego, and J. Calvo-Zaragoza, "Automatic detection of inconsistencies between numerical scores and textual feedback in peer-assessment processes with machine learning," *Computers & Education*, vol. 140, p. 103609, 2019, doi: https://doi.org/10.1016/j.compedu.2019.103609.

[23] A. E. Waters, D. Tinapple, and R. G. Baraniuk, "BayesRank: A Bayesian Approach to Ranked Peer Grading," in *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, in L@S '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 177–183. doi: 10.1145/2724660.2724672.

[24] Y. Zhang and E. F. Gehringer, "Can Students Produce Effective Training Data to Improve Formative Feedback?," in *2021 IEEE Frontiers in Education Conference (FIE)*, 2021, pp. 1–7. doi: 10.1109/FIE49875.2021.9637414.

[25] W. Hart-Davidson, R. Omizo, and M. Meeks, "Detecting High-Quality Comments in Written Feedback with a Zero Shot Classifier," in *Proceedings of the 39th ACM International Conference on Design of Communication*, in SIGDOC '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 319–325. doi: 10.1145/3472714.3473659.

[26] Y. Xiao, Y. Gao, C. Yue, and E. Gehringer, "Estimating Student Grades through Peer Assessment as a Crowdsourcing Calibration Problem," in *2022 20th International*

Conference on Information Technology Based Higher Education and Training (ITHET), 2022, pp. 1–9. doi: 10.1109/ITHET56107.2022.10031993.

[27] M. Selmi, H. Hage, and E. Aïmeur, "Evaluating LSA sensibility to disclosure in learners' interactions," in *2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA)*, 2015, pp. 1–8. doi: 10.1109/SITA.2015.7358384.

[28] Y. Han, W. Wu, Y. Yan, and L. Zhang, "Human-Machine Hybrid Peer Grading in SPOCs," *IEEE Access*, vol. 8, pp. 220922–220934, 2020, doi: 10.1109/ACCESS.2020.3043291.

[29] J. Xu, J. Liu, P. Lv, and P. Yang, "Improving Peer Assessment Accuracy by Incorporating Grading Behaviors," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2021, pp. 1162–1169. doi: 10.1109/ICTAI52525.2021.00184.

[30] O. Luaces, J. Díez, A. Alonso, A. Troncoso, and A. Bahamonde, "Including Content-Based Methods in Peer-Assessment of Open-Response Questions," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015, pp. 273–279. doi: 10.1109/ICDMW.2015.256.

[31] C. Cachero, J. R. Rico-Juan, and H. Macià, "Influence of personality and modality on peer assessment evaluation perceptions using Machine Learning techniques," *Expert Systems with Applications*, vol. 213, p. 119150, 2023, doi: 10.1016/j.eswa.2022.119150.

[32] X. Fan, Y. Liu, N. Cao, J. Hong, and J. Wang, "MindMiner: Quantifying Entity Similarity via Interactive Distance Metric Learning," in *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, in IUI Companion '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 93–96. doi: 10.1145/2732158.2732173.

[33] Y. Xiao *et al.*, "Modeling review helpfulness with augmented transformer neural networks," in *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, 2022, pp. 83–90. doi: 10.1109/ICSC52841.2022.00019.

[34] V. Bolón-Canedo, J. Díez, O. Luaces, A. Bahamonde, and A. Alonso-Betanzos, "Paving the way for providing teaching feedback in automatic evaluation of open response assignments," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 3447–3453. doi: 10.1109/IJCNN.2017.7966289.

[35] Z. Fan, M. LU, and X. Li, "Peer Assessment Based on the User Preference Matrix," in *2020 International Conference on Artificial Intelligence and Education (ICAIE)*, 2020, pp. 1–4. doi: 10.1109/ICAIE50891.2020.00009.

[36] M. Parvez Rashid, E. F. Gehringer, M. Young, D. Doshi, Q. Jia, and Y. Xiao, "Peer Assessment Rubric Analyzer: An NLP approach to analyzing rubric items for better peer-review," in *2021 19th International Conference on Information Technology Based Higher Education and Training (ITHET)*, 2021, pp. 1–9. doi: 10.1109/ITHET50392.2021.9759679.

[37] M. S. M. Sajjadi, M. Alamgir, and U. von Luxburg, "Peer Grading in a Course on Algorithms and Data Structures: Machine Learning Algorithms do not Improve over Simple Baselines," in *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, in L@S '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 369–378. doi: 10.1145/2876034.2876036.

[38] C. E. Kulkarni, R. Socher, M. S. Bernstein, and S. R. Klemmer, "Scaling short-answer grading by combining peer assessment with algorithmic scoring," in *Proceedings of the First ACM Conference on Learning @ Scale Conference*, in L@S '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 99–108. doi: 10.1145/2556325.2566238.

[39] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, I. Estévez-Ayres, and C. D. Kloos, "Sentiment analysis in MOOCs: A case study," in *2018 IEEE Global Engineering Education Conference (EDUCON)*, 2018, pp. 1489–1496. doi: 10.1109/EDUCON.2018.8363409.

[40] F. Sciarrone and M. Temperini, "Simulating Massive Open On-line Courses Dynamics," in *2019 18th International Conference on Information Technology Based Higher Education and Training (ITHET)*, 2019, pp. 1–9. doi: 10.1109/ITHET46829.2019.8937336.

[41] A. V. Y. Lee, "Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation," *Studies in Educational Evaluation*, vol. 77, p. 101250, 2023, doi: 10.1016/j.stueduc.2023.101250.

[42] E. Scheihing, M. Vernier, J. Guerra, J. Born, and L. Crcamo, "Understanding the Role of Micro-Blogging in B-Learning Activities: Kelluwen Experiences in Chilean Public Schools," *IEEE Transactions on Learning Technologies*, vol. 11, no. 3, pp. 280–293, 2018, doi: 10.1109/TLT.2017.2714163.

[43] A. Bürgermeister, I. Glogger-Frey, and H. Saalbach, "Supporting Peer Feedback on Learning Strategies: Effects on Self-Efficacy and Feedback Quality," *Psychology Learning & Teaching*, vol. 20, pp. 383–404, 2021, doi: 10.1177/14757257211016604.

[44] A. Sizo, A. Lino, Á. Rocha, and L. P. Reis, "Defining quality in peer review reports: a scoping review," *Knowledge and Information Systems*, vol. 67, pp. 6413–6460, 2025, doi: 10.1007/s10115-025-02435-0.

[45] H. Baer, E. Legome, D. Satnick, J. McHugh, and G. Loo, "A New Teaching Tool for Peer Review of Charting and Care in the Emergency Department," *The Joint Commission Journal on Quality and Patient Safety*, vol. 49, no. 2, pp. 105–110, 2023, doi: 10.1016/j.jcjq.2022.10.007.

[46] C. Rastogi *et al.*, "A randomized controlled trial on anonymizing reviewers to each other in peer review discussions," *PLoS ONE*, vol. 19, no. 12, 2024, doi: 10.1371/journal.pone.0315674.

[47] J. L. Hill, K. Berlin, J. Choate, L. Cravens-Brown, L. McKendrick-Calder, and S. F. Smith, "Exploring the Emotional Responses of Undergraduate Students to Assessment Feedback: Implications for Instructors," *Teaching & Learning Inquiry*, vol. 9, no. 1, 2021, doi: 10.20343/teachlearninqu.9.1.20.

[48] O. Akbari and J. Sahibzada, "Students' Self-Confidence and Its Impacts on Their Learning Process," *American International Journal of Social Science Research*, vol. 5, no. 1, pp. 1–15, Jan. 2020, doi: 10.46281/aijssr.v5i1.462.

[49] A. Sunar and M. S. Khalid, "Natural Language Processing of Student's Feedback to Instructors: A Systematic Review," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 741–753, 2024, doi: 10.1109/tlt.2023.3330531.

[50] S. Ibragimova *et al.*, "A Longitudinal Study on NLP-Enhanced Bilingual Pedagogy for Non-Linguistic Majors," *Digital Technologies Research and Applications*, 2026, doi: 10.54963/dtra.v5i1.1985.