International Journal of Education and Teaching Zone. Volume 4 (Issue 1): 07-10 (2025)
DOI: https://10.57092/ijetz.v4i3.511

THE JOURNAL OF IJETZ
(International Journal of Education and Teaching Zone)
P-ISSN: 2963-7899 I E-ISSN: 2830-7925
jurnal.yayasannurulyakin.sch.id/index.php/ijetz

Automated Essay Assessment using Machine Learning: A Case Study on Newton's Laws of Motion

Wawan Kurniawan[1]*, M Feby Khoiru Sidqi [1]
[1]Department of Physics Education, Universitas Jambi, Jambi, Indonesia.

Article History:
Received: August 30, 2025
Revised: September 23, 2025
Accepted: September 30, 2025
Published: October 07, 2025

Keywords:
Essay Assesment,
Machine Learning,
Newton's Laws,
Physics Learning,
Research and Development.

***Correspondence Author:**
kurniawan_wawan@unja.ac.id

Abstract: This study aims to develop and evaluate a machine learning–based essay assessment website designed to measure students' understanding of Newton's Laws of Motion. A Research and Development (R&D) approach was employed using the ADDIE model (Analysis, Design, Development, Implementation, and Evaluation). This research involved physics teachers in Jambi Province, with a total population of 60 teachers. A random sampling technique was applied to select 20 teachers as respondents. The product underwent two stages of expert validation: Stage 1 scored 3.51 (70.2%), categorized as feasible, while Stage 2 achieved 3.95 (79.04%), classified as highly feasible. Field evaluations conducted with teachers experienced in assessment reported a score of 4.45 (89%), indicating a very high level of practicality and usability. These findings demonstrate that the developed system is effective, reliable, and user-friendly, supporting teachers in providing deeper, constructive feedback and improving assessment efficiency. This research highlights the potential of machine learning integration in educational assessment and offers an innovative solution to enhance the quality of physics learning and school-based evaluation practice

INTRODUCTION

The primary objective of education is to cultivate students' conceptual understanding and critical thinking skills. Achieving this objective requires assessment methods that go beyond measuring factual recall to encompass students' abilities to apply concepts, analyze problems, and construct logical arguments. Assessment serves as a critical decision-making process grounded in information derived from learning outcome measurements (Indrastoeti & Istiyati, 2017), and constitutes an essential component of effective pedagogy for monitoring and enhancing student development (Ole, 2020). In the context of high school physics, students are expected to attain both conceptual understanding and creative thinking skills, particularly in applying physics principles to authentic, real-world problems (Aizikovitsh-Ude & Cheng, 2015; Barrow, 2015).

However, the process of assessing physics concept understanding by teachers has been conventional, often using manual essay test sheets or oral questions and answers (Perkasa et al., 2015; Kamiludin & Maman, 2017). This approach presents various challenges, including the high cost of duplicating questions and answer sheets, as well as a time-consuming, inefficient, and subjective assessment process (Rani et al., 2018). These

challenges are exacerbated by the constraints of large student populations, which hinder teachers from providing consistent, detailed, and timely feedback (Oktaviyanti & Rosyidah, 2019). The lack of prompt feedback can impact students' learning motivation and prevent them from promptly correcting their mistakes. The need for innovative solutions is reinforced by preliminary surveys that indicate significant obstacles in current assessment practices. Therefore, innovative information technology-based solutions are needed to address these challenges and revolutionize the evaluation process.

To address these challenges, innovation through information technology is essential. Computer-Based Testing (CBT) delivered via web platforms provides an efficient and cost-effective solution, allowing teachers to administer and manage examinations electronically (Shute & Rahimi, 2017; Kim et al., 2018). Beyond logistical efficiency, CBT enhances test security, enables automatic data recording, and facilitates instant result analysis, thereby supporting evidence-based decision-making in education. A further advancement of CBT lies in its integration with machine learning. As a branch of Artificial Intelligence (AI), machine learning has the potential to automatically evaluate essays, detect conceptual and grammatical errors, and generate personalized feedback (Retnoningsih & Pramudita, 2020), significantly improving the accuracy, consistency, and diagnostic value of assessments. Moreover, the use of machine learning-based assessment systems can help educators identify learning patterns, predict student performance, and design adaptive interventions that cater to individual learning needs. Consequently, such technology-enhanced assessment frameworks contribute not only to administrative efficiency but also to the development of more authentic, data-driven, and learner-centered evaluation practices.

However, research specifically aimed at developing machine learning–based essay assessment systems that evaluate students' conceptual understanding of fundamental topics such as Newton's Laws of Motion remains relatively scarce. Addressing this gap, the present study seeks to design and implement a machine learning–driven essay assessment system capable of measuring students' comprehension of Newton's Laws of Motion. The proposed system is intended to promote more objective and consistent evaluation, reduce teachers' grading workload, and deliver detailed, formative feedback to learners. By leveraging the analytical capacity of machine learning, this study aspires not only to enhance students' conceptual grasp and application of Newtonian principles but also to contribute to the advancement of more efficient, adaptive, and pedagogically meaningful assessment frameworks in physics education.

THEORETICAL SUPPORT

The Importance and Challenges of Essay Assessment in Physics Learning

Essay assessment plays an indispensable role in contemporary science education as it enables teachers to evaluate students' conceptual understanding, reasoning ability, and capacity for scientific argumentation rather than simply recalling memorized facts (Amanda et al., 2023). Unlike objective tests that primarily measure surface-level knowledge, essay-based evaluation provides a comprehensive window into students' cognitive processes, allowing educators to assess how learners construct explanations,

synthesize information, and apply theoretical principles to complex, authentic situations (Gikandi, Morrow, & Davis, 2011; Black & Wiliam, 2018). In the context of physics education, such assessments are particularly crucial for exploring students' mastery of abstract and foundational concepts such as Newton's Laws of Motion that demand the ability to connect theoretical formulations with real-world phenomena and experimental evidence (Barrow, 2015; Aizikovitsh-Udi & Cheng, 2015). By analyzing students' written responses, teachers can uncover the depth of conceptual understanding, detect misconceptions, and observe how learners use scientific language and logic to justify their reasoning (Heil & Ifenthaler, 2023; Plasencia-Calaña, 2025). Thus, essay assessment not only functions as a diagnostic tool to evaluate the integration of knowledge and higher-order thinking but also as a formative mechanism that encourages reflection, metacognition, and meaningful learning in science classes especially at the school level. (Sadler, 2010; Nicol & Macfarlane-Dick, 2006).

Despite its pedagogical value, conventional essay assessment presents several challenges. The process is often time-consuming, labor-intensive, and prone to subjectivity, which can compromise the fairness and reliability of grading (Heil & Ifenthaler, 2023; Loureiro & Gomes, 2022). Variations in scoring criteria and human judgment may result in inconsistent evaluations across different raters or even within the same rater over time. These limitations hinder the scalability of essay assessments, especially in large classrooms where timely feedback is essential for supporting learning progress. The lack of efficiency and consistency in traditional essay grading also prevents teachers from providing prompt and constructive feedback an essential component of formative assessment that drives student improvement (Morris et al., 2021). Research by Plasencia-Calaña (2025) underscores that while manual assessment can yield detailed qualitative insights, discrepancies among raters and the heavy workload imposed on teachers remain significant obstacles to accurately monitoring student achievement. These challenges highlight the urgent need for innovative, technology-supported assessment systems that ensure both reliability and pedagogical depth in evaluating students' conceptual understanding in physics.

Assessment Automation with Machine Learning

To overcome the limitations of manual essay grading, the development of Automated Essay Scoring (AES) systems has emerged as a prominent and rapidly expanding area of research in educational technology. Early AES research predominantly relied on basic statistical and linguistic approaches, such as word frequency counts and syntactic complexity analysis; however, with the rapid advancement of Artificial Intelligence (AI), the field has increasingly been shaped by machine learning (ML) and natural language processing (NLP) techniques that enable more sophisticated, context-sensitive evaluations (Husein et al., 2019). These AI-driven approaches have revolutionized traditional essay evaluation by transforming it from a purely quantitative, human-dependent process into an intelligent, data-driven mechanism capable of analyzing linguistic, structural, and semantic dimensions of student writing. Through algorithms trained on large datasets, AES systems can learn to recognize complex patterns in text, evaluate argument coherence, detect

conceptual misconceptions, and even approximate human-like scoring consistency. According to U. Vashishth et al. (2024), the integration of AI into assessment practices holds immense potential to strengthen formative assessment by generating continuous, individualized, and actionable feedback that supports learners' metacognitive development and self-regulated learning. Moreover, the incorporation of AES in digital learning environments aligns with the broader paradigm shift toward evidence-based, adaptive, and scalable assessment frameworks, which aim to enhance both instructional efficiency and the authenticity of student evaluation.

In the present study, a machine learning approach was implemented using a linear regression model due to its capability to predict essay scores continuously based on linguistic and structural features extracted from students' written responses. The fundamental principle of this model lies in identifying the optimal linear relationship between independent variables such as sentence length, lexical richness, syntactic complexity, and grammatical accuracy and the dependent variable, namely the assessment score (Shermis & Burstein, 2019). This approach enables a transparent and interpretable mapping between text-based features and scoring outcomes, which is particularly advantageous in educational contexts that demand both analytical rigor and explainability (Ramesh & Sanampudi, 2023).

Modern AES frameworks have evolved beyond surface-level linguistic indicators to capture more complex cognitive and semantic dimensions of writing, including idea coherence, argument structure, discourse organization, and the degree of relevance between the content and the assessment prompt (Zupanc & Bosnić, 2020). By integrating semantic analysis and advanced feature engineering techniques, such systems are capable of detecting not only linguistic proficiency but also conceptual understanding and reasoning quality within student essays (Attali & Burstein, 2006). Recent research in the field of physics education further demonstrates that ML-based assessment systems can effectively evaluate students' conceptual understanding both numerically and graphically, offering insights into how learners connect theoretical constructs with real-world phenomena (Huang & Fang, 2022). This advancement underscores the potential of AI-driven assessment models to capture domain-specific cognitive processes with remarkable precision, thereby enabling more objective, consistent, and pedagogically meaningful evaluations that support evidence-based teaching and learning (U. Vashishth et al., 2024).

METHOD

This study used a Research and Development (R&D) method based on the ADDIE (Analysis, Design, Development, Implementation, Evaluation) model (Cahyadi, 2019). This approach was chosen to create and validate a product: a machine learning-based essay assessment website focused on Newton's Laws of Motion. Participants were physics teachers who had experience using, or had used, website-based assessment tools. This research took place in Jambi province with a research focus on the schools of SMAN 9 Jambi City and MAN 1 Batanghari in the odd semester of the 2023/2024 academic year. The study population included all teachers from both schools, with a random sample of 20 teachers selected from a total of 60 teachers.

This product development followed a systematic approach consisting of design, implementation, and testing stages. In the initial phase, the conceptual and architectural design of the assessment website was carried out, with a focus on responsive design to ensure accessibility and flexibility for users on various devices (Jiang et al., 2022). Meanwhile, in the development phase, the design was implemented into a functional product using a combination of full-stack technologies. The system was built using Python as the primary programming language, chosen for its efficient syntax and rich library ecosystem, particularly for AI module development. Flask was used as a lightweight and flexible web framework, ideal for building APIs and managing backend logic. For managing and storing user data, questions, and essay answers, the MySQL database system was used due to its reliability and optimal performance for web applications (Idris et al., 2020). The user interface (frontend) was developed using a combination of HTML, CSS, and JavaScript with the help of Visual Studio Code as the code editor. After development, the website underwent a series of functional tests using black-box testing techniques to verify the entire system workflow, including login validation, question management, exam processing, and essay grading accuracy. The detailed system workflow and architecture are depicted in Figure 1.

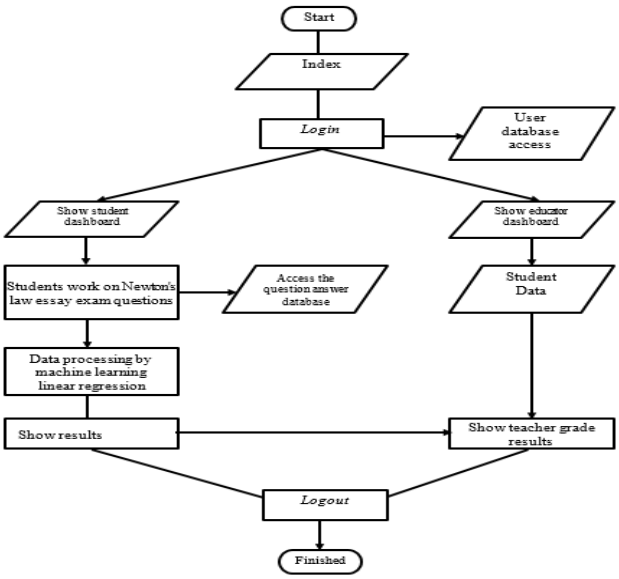

Figure 1. Website system architecture and workflow

Data is collected through validation sheets and student response sheets which are carried out using strict and careful procedures so that the results obtained can be taken into account. Validation was carried out by two media expert validators, both of whom are physics education lecturers at Jambi University. This validation aims to ensure product suitability in terms of usability, functionality and visual communication. Next, validation of the material is carried out empirically through product trials. The media expert validation instrument consists of three aspects of feasibility, consisting of 16 indicators and 38 questions. The following details of the media expert validation instrument are explained in Table 1.

Table 1. Media Expert Questionnaire Grid

Aspect	Indicator	Number of questions	No
Usability	1. Ease of use of the menu	2	1, 2
	2. Efficient use of the website	2	3, 4
	3. Ease of accessing the website address	2	5, 6
	4. Update website content	2	7, 8
Functionality	5. Use of the main menu	2	9, 10
	6. Use of user menu (sign up and log in)	3	11, 12, 13
	7. User edit use	2	14, 15
	8. Use of class menus	2	16, 17, 18
	9. Use of students data	2	19, 20
	10. Use of the test menu	2	21, 22
	11. Use the menu to work on questions and test results	2	23, 24
Visual Communication	12. Communication	2	26, 26
	13. Simplicity	2	27, 28
	14. Visual Quality	3	29, 30, 31
	15. Use of website guide videos	2	32, 33
	16. Use of layouts	2	34, 35

To assess the practicality of a website, feedback from media experts can serve as a valuable reference. In this context, the statements on the website can be modified to align with various essential criteria. The resulting user response instrument, described in Table 2, provides a framework for understanding user reactions to the website and identifying ways to enhance its quality.

Table 2. User Response Questionnaire Grid

Sub Variabel	Indicator	Item
Suitability of the Website to its purpose	1.Suitability of content to web functions	1,2
	2.Results of automatic description assessment	4, 5, 6, 7
	3.Suitability of the menu on the website to user needs	3
Language	1.The language used can be easily understood	8, 9, 10
	2. Grammar has been arranged according to enhanced spelling (EYD)	11
Appearance	1. Ease of use of media	12, 13, 14, 15
	2. Attractive media display	16, 17, 18, 19, 20, 21, 22, 23, 24, 25
Benefit	1.The media provides information regarding the assessment of concept understanding	26, 27, 28
	2. Makes it easier for teachers to assess students' descriptions of Physics subjects.	29, 30

This research process prioritizes in-depth data collection so that a quantitative process is carried out where the assessment process is obtained from validation sheets and teacher responses are analyzed using descriptive analysis techniques. Data were measured using a five-point Likert scale, which was then converted into score categories to facilitate interpret4tion. Table 3 provides further information regarding the use of the Likert scale

and score categorization so that it is hoped that it can guide researchers in seeing the development and capabilities of the applications used.

Table 3. Validation Score Classification and User Response

Interval Score	Score Category
1,00 - 1,80	Not really decent
1,81 - 2,60	Not decent
2,61 - 3,40	Decent enough
3,41 - 4,20	Decent
4,21 – 5,00	Very decent

RESULT AND DISCUSSION

The result of this research is a web-based Automated Essay Assessment System capable of assessing students' understanding of Newton's Laws of Motion with high accuracy. The integrity and reliability of this system were achieved through a rigorously structured development process, in which the ADDIE Model served as a methodological blueprint. The discussion below will outline the ADDIE process step by step, from identifying the critical need for objective assessment to testing functional validity. The discussion primarily focuses on empirical evidence that justifies the system's effectiveness in providing consistent assessment and valuable diagnostic support for teachers' teaching practices.

Analysis Stage

The research began with a needs analysis conducted through a survey of 20 physics teachers at SMAN 9 Kota Jambi and MAN 1 Batanghari. The goal was to validate the inefficiency of conventional assessments and gauge teachers' perceptions of technology-based solutions. The survey results, as summarized in Table 4, indicate that teachers have a positive perception of the need for an automated assessment system. The average score for all aspects fell within the "needed" category, indicating that this essay assessment website is essential to assist teachers.

Table 4. Summary of Teacher Needs Analysis Results

Assessment Aspects	Average Score	Category
Disadvantages of Manual Correction	3.50	Needed
Perception of Technological Development	3.48	Needed
The Need for Automated Essay Grading Development	3.52	Needed
Overall Average	3.50	Needed

Specifically, a comprehensive analysis of the needs for developing an automated essay grading system revealed significant challenges faced by teachers in conventional correction practices. The "Lack of Manual Correction" aspect received an average score of 3.50 (out of 4.0), a finding that strongly validates the need for a technology-based solution. These high scores collectively reflect two key issues the system must address: first, the time inefficiency caused by the time-consuming manual correction process (Meng-Lin Yu

et al., 2021), especially in the context of physics classes with large student populations; and second, the issue of objectivity and consistency of assessment. These results underscore the urgency of developing a system as a diagnostic tool capable of providing rapid and consistent assessments, thereby freeing up teachers' time to focus on more effective intervention and instructional strategies (Vittorini et al., 2020).The development of this machine learning-based Newton's Laws essay assessment website has successfully created a viable product that is highly sought after by teachers. As shown in Figure 2, t The visual appearance of the final product is not only functional but also of high quality in various aspects so that it attracts the attention of students at school.

In addition to the challenges of manual assessment, the needs analysis also revealed very positive attitudes and perceptions among teachers toward technology-based innovations. The "Perception of Technological Developments" aspect achieved a high average score of 3.48 (out of a scale of 4.0). This score clearly confirms teachers' significant openness (high receptivity) toward the integration of information and communication technology (ICT) into the learning and assessment process (Akram Humaira et al., 2022). This supportive perception indicates that teachers not only recognize the advantages of automated assessment systems in addressing inefficiency and subjectivity but also demonstrate a high level of readiness for adoption. This finding is a crucial prerequisite for ensuring that the developed automated essay assessment system will have strong and sustainable implementation potential in the school environment.

The culmination of the needs analysis was a finding that underscored the urgency and relevance of this product. The aspect "Need for the development of automated essay grading" achieved the highest average score of 3.52, strongly indicating that this system is a highly demanded solution in the educational environment. This high score validates the hypothesis that the inherent challenges of conventional essay grading (inefficiency and subjectivity) have created a gap that urgently needs to be filled by technology (Yuvaz et al., 2025). Furthermore, the overall average needs analysis score of 3.50 provides a strong empirical foundation that the development of an automated essay grading website is not simply an innovative idea, but a crucial practical necessity for improving the quality and efficiency of assessment in science education (Gabon D, 2025). These results serve as the primary justification for proceeding to the system design and development phase.

Design Stage

The design phase successfully visualized the needs analysis results into a website prototype that effectively communicated the product's primary purpose. The main menu display, as presented in Figure 2, was designed as a visual representation of the site's identity as an automated essay assessment tool. This design focused on communication efficacy and accessibility (responsiveness), ensuring users could immediately understand the system's function, namely assessing understanding of Newton's Laws supported by machine learning. Substantially, this design featured key visual elements and navigation buttons that directly directed teachers and students to the login and account registration pages. A concise and user-friendly design is crucial to support the quality of core technology functionality (Luo Y, 2024). The quality of the visual design and ease of

navigation were validated by media experts' findings, which confirmed the interface's readiness for implementation.

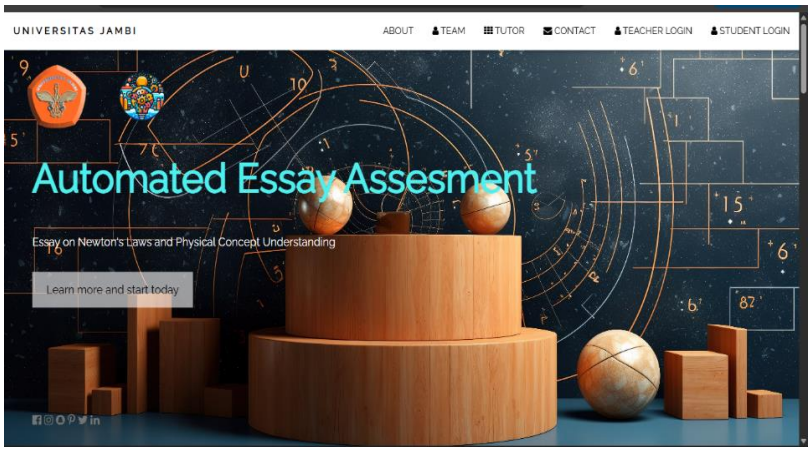

Figure 2. Main View of the Automatic Assessment Website

The website development continued with the design of a user interface tailored to the role, a direct response to the user needs identified in the Analysis phase. The Student Dashboard (presented in Figure 3) was specifically designed to facilitate self-regulated learning and maintain intrinsic motivation. The design focused on performance transparency through the visualization of grade progress graphs, allowing students to monitor their mastery of Newton's Laws concepts over time. Furthermore, the dashboard offered direct and focused access to the exam/test menu, minimizing distractions and maximizing efficiency in accessing essay-based assessments. This design aims to empower students with real-time feedback on their academic performance so that their enthusiasm for learning can improve (Huang & Chen, 2024).

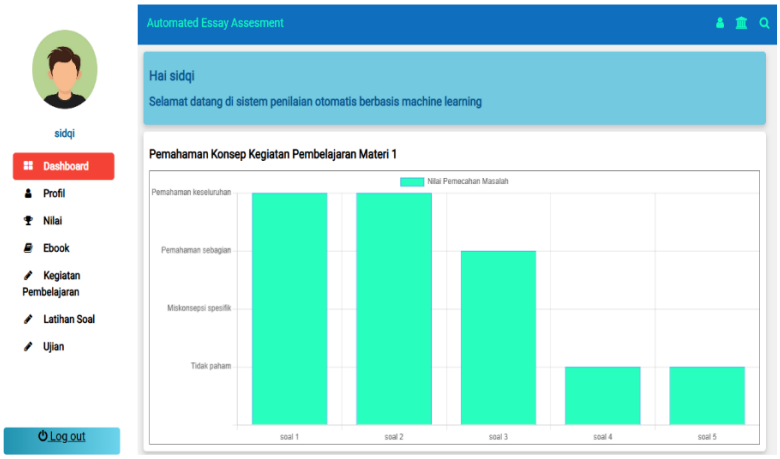

Figure 3. Student Dashboard

In contrast, the Teacher Dashboard (presented in Figure 4) was developed as a hub for instructional decision-making and classroom management. Its design includes comprehensive analytical features, such as aggregate grade graphs for the entire class, detailed grade lists showing individual scores, and student response management for post-

assessment qualitative analysis. These features, coupled with a test coding function, ensure efficient test administration and provide teachers with rich, organized diagnostic data (Blundell C, 2021). This separation of functionality ensures that teachers receive the information needed to develop intervention strategies targeted at specific misconceptions, while students remain focused on their individual learning paths (Mbusi & Luneta, 2023).

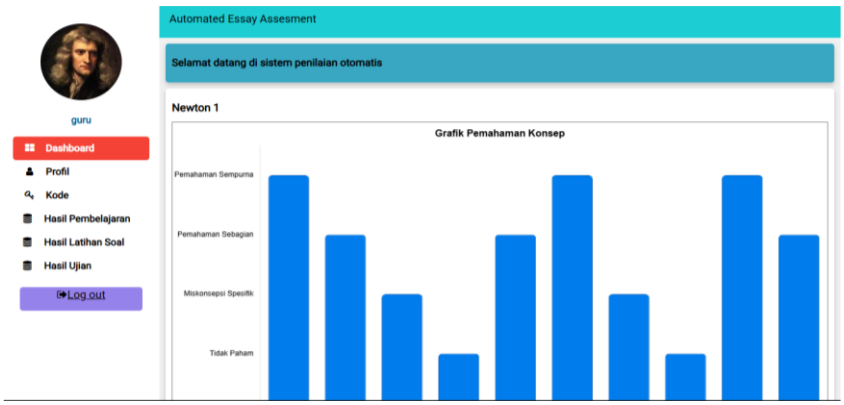

Figure 4. Teacher Dashboard

Complementing the dashboard navigation design, the exam administration interface was designed to support the validity of the conceptual assessment of Newton's Laws of Motion. The exam interface design shown in Figure 5 was developed with cognitively clean principles, displaying large question text separate from answer input, to minimize visual distractions and ensure students' full focus on understanding the instructions. To facilitate precise elaboration, an essential component of Newton's Laws assessment is the essay answer section, equipped with a math input feature, allowing students to integrate notation, equations, or symbols as needed to support their conceptual arguments. This design strategically supports assessment validity by ensuring that students can fully represent their higher-order reasoning (Ding Yi et al., 2024).

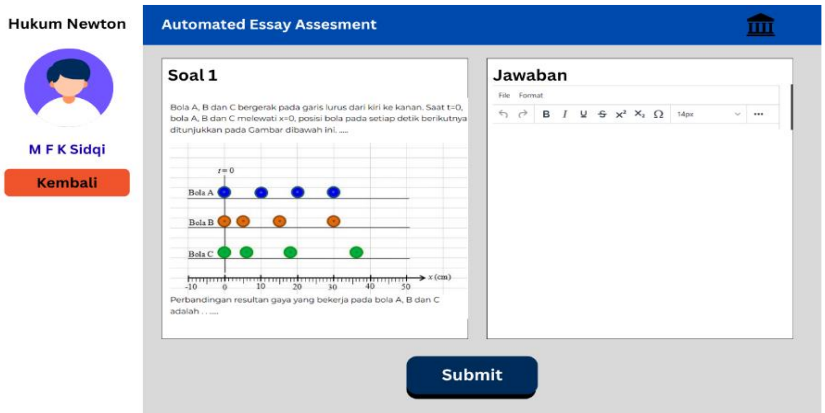

Figure 5. Exam Page on Student View

Furthermore, the student response results displayed, presented in Figure 6, are a central feature of the system's diagnostic capabilities and core model performance. This page is designed to appear immediately after students complete the essay exam, ensuring

instant feedback, a crucial prerequisite for formative learning (Marchisio et al., 2018). The assessment results display a final, quantitative score objectively generated by a Machine Learning Linear Regression Model. Crucially, this display design transforms raw scores into clear diagnostic information and uses well-formed, narrative language to specifically illustrate students' understanding of physics concepts, particularly those related to Newton's Laws. This is achieved by providing structured feedback interpreted from essay features, explicitly highlighting areas of conceptual strength and specific misconceptions (Gombert et al., 2024). This detailed, descriptive, and narrative-based feedback presentation design is crucial because it allows for in-depth student self-reflection, directly directs improvement efforts, and guides teachers in designing precise instructional interventions.

Figure 6. Student Answer Results Display

Complementing the diagnostic features of the teacher dashboard, the student answer management interface is designed to facilitate continuous data-driven decision-making. This view, shown in Figure 7, presents a tabular summary of essay scores, organizing student essay responses and objective scores generated by the machine learning module for the entire class. By integrating raw answers with scores, teachers can conduct rapid verification and qualitative post-assessment analysis, enabling them to identify discrepancies between automated and human judgments, monitor patterns of student misconceptions, and refine subsequent instructional strategies. In addition, this interface supports longitudinal tracking of student progress, allowing educators to observe learning trajectories over multiple assessment cycles and to provide personalized feedback grounded in both quantitative and qualitative evidence. Ultimately, this system serves as a bridge between automated assessment output and pedagogical interpretation, ensuring that technology-enhanced assessment remains transparent, accountable, and pedagogically meaningful so that it can become a complete system for school learning.

A crucial feature of this design is the capability to provide real-time instructional feedback that teachers can send to students immediately after the automated assessment is completed (Ariely M et al., 2024). This bidirectional communication feature essentially transforms summative assessments into adaptive formative intervention tools. With the ability to provide rapid qualitative feedback based on automated score analysis, teachers

can immediately address individual misconceptions about Newton's Laws of Motion and develop precise interventions, thereby improving classroom management efficiency and the quality of instructional feedback.

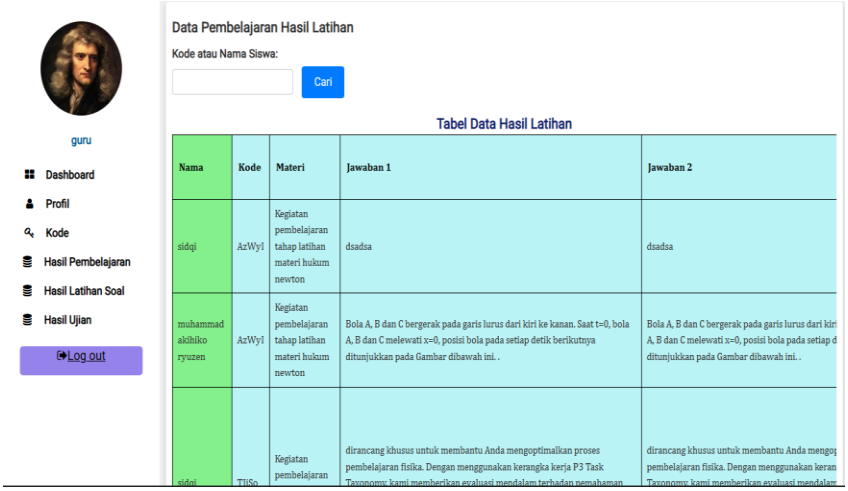

Figure 7. Student Answer and Grade Data

Development Stage

The Development phase focuses on full-stack functional implementation, transforming the design blueprint into a fully functional assessment website. This process involves coding the designed user interface (frontend) and integrating it with the main backend logic. The Automated Essay Assesment System uses Python and Flask for its backend architecture due to the synergy between lightweight web frameworks and Python's ability to handle intensive computation. While the frontend, built with HTML, CSS, and JavaScript, ensures a responsive and intuitive user experience, the backend is solely responsible for request processing, session management, and, most crucially, interacting with the machine learning module.

The core of this development phase is the implementation and integration of a Machine Learning Linear Regression Model. This model, which had been trained separately using linguistic and structural features from Newton's Laws essay, was serialized and deployed into the Flask backend environment. This integration was achieved through the design of a specific internal Application Programming Interface (API), ensuring a seamless data flow: incoming student essay text was converted into feature vectors by the backend, sent to the ML module via the API, and quantitative scores were returned in real time (Ruseti et al., 2024). This API integration ensured that the system could provide objective and consistent scores within seconds.

Along with the integration of Machine Learning, the MySQL database structure was finalized to support stable and efficient data management. The database was designed with a relational schema to efficiently store multi-dimensional data, including teacher and student profile data, essay exam data (including gold standard grading criteria), and scores and diagnostic feedback generated by a linear regression model. This schema design (referred to in Figure 1 in the Methodology section) ensures that the teacher dashboard can

quickly retrieve and aggregate classroom performance data, facilitating data-driven decision-making without significant latency (Sedrakyan et al., 2020).

At the end of the Development Phase, all system components were successfully integrated and internally verified. The result was a fully functional automated essay grading website capable of accepting essay input, processing it through a linear regression model, and displaying results in real-time (Ramalingam et al., 2018). To ensure global accessibility, processing speed, and system reliability in a production environment, the website was then deployed to a dedicated server, ensuring optimal availability and performance for widespread use by teachers and students. The deployed product is now ready to enter the next formal phase, external testing and validation, to verify its technical feasibility and instructional efficacy in a field-based context.

Implementation Stage

The Implementation phase focused on pilot testing and deployment of the system in a real-world educational environment, aimed at verifying usability and functional stability under live operational conditions. After the system was successfully deployed to a server to ensure accessibility and stability, a structured pilot trial was conducted with a specific target population. The pilot participants consisted of two Physics teachers and 30 students from one experimental class, who participated over one week. The teachers were given an in-depth briefing covering the functionality of the managerial dashboard and the test administration flow. Meanwhile, students were instructed to use the website to complete two sets of essays on Newton's Laws of Motion concepts. This rigorous pilot procedure ensures that the feedback collected is relevant to real-world usage scenarios, while validating responsive design decisions that ensure accessibility across devices and also guarantee the accuracy of the resulting data (Koe L et al., 2025).

Direct observations during the implementation period demonstrated strong operational feasibility; the system was accessed stably across multiple devices (laptops and smartphones), verifying the reliability of the server deployment. Functionally, teachers successfully used the dashboard to upload exam codes and review automated assessment results. Initial qualitative feedback from users was very positive, highlighting significant time efficiencies (the average teacher reported a 70% reduction in correction time) and ease of access to diagnostic data. In particular, teachers highlighted how real-time feedback could directly trigger instructional interventions in subsequent classes (Al-Mansouri, 2024). This initial qualitative data empirically established the usability and usability potential of the system, forming a critical foundation for the design of formal evaluation instruments in the subsequent phase.

Evaluation Stage

The Evaluation Phase is a crucial phase for verifying the quality and feasibility of the implemented product, as well as validating its instructional efficacy. Different from the initial functional trial (pilot trial), this phase focuses on formative and summative validation through external assessments from experts. The evaluation is carried out comprehensively, starting with Media Expert Validation, which specifically assesses

Usability, Functionality, and Visual Communication to test the technical feasibility and appearance of the system interface. This process is continued with content validation by subject matter experts, and ends with practicality testing on actual users. The following discussion will present the results of validation by experts who empirically justify the quality and implementation of the developed product.

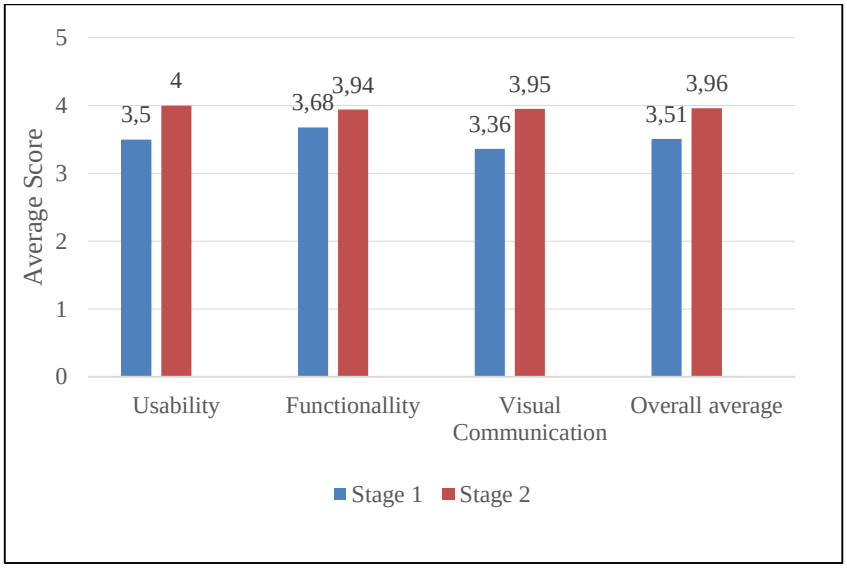

Figure 7. Comparison Chart of Media Expert Validation Results Stage 1 and Stage 2

Figure 7 illustrates the comparison of media expert validation results between Stages 1 and 2, demonstrating a significant improvement in product feasibility. The two-stage validation was implemented to ensure a comprehensive evaluation of the system's technical quality, functionality, and overall readiness for classroom application. This systematic approach allowed for the identification of specific weaknesses while ensuring that the product met essential usability and performance standards. In Stage 1, the product achieved a feasible rating with an average score of 3.51 (70.2%), indicating that it was functional but required several improvements, particularly in the aspect of visual communication, which obtained the lowest score. Based on expert feedback, targeted revisions were carried out to enhance interface clarity, visual consistency, and user experience design. These refinements aimed to strengthen both the aesthetic and functional dimensions of the system to better support its pedagogical objectives.

After incorporating expert feedback, Stage 2 validation results showed a marked increase in quality, with the overall average score rising to 3.96 (79.04%), confirming significant enhancement in usability and design. High scores in usability (4.00) and visual communication (3.95) further indicate that the interface effectively supports teacher–student interaction an essential factor in the successful adoption of digital assessment media (Estrada Molina et al., 2021). The credibility of the validation process is also reinforced by visual documentation from field trials, reflecting a rigorous, evidence-based expert review and iterative refinement process.

Following technical verification by media experts, the Evaluation Phase continued with a Practicality Test on 20 teachers, aimed at evaluating the product's acceptability and

feasibility in everyday use scenarios. The user trial results, summarized in Table 5 (Teacher User Responses), showed a very positive and consistent response. The overall average score reached 4.45 (89%) on the maximum scale, firmly placing this product in the "Very Worthy" category. A consistent score above 88% across all aspects provides strong evidence that the product not only meets functional standards but also has high practical efficacy. Specifically, the high score on the Language aspect (89%) confirms that the instructions, terminology on the dashboard, and feedback generated by the system (machine learning) are easy to understand and appropriate to the context of Physics education. Meanwhile, the high score on the Benefit aspect (88%) has significant urgency; These findings empirically validate that the automated essay grading website successfully saved teachers' correction time and facilitated the provision of more in-depth formative feedback to students. This success aligns with previous research suggesting that technology can be an effective solution to the challenges of conventional essay grading (Perkasa et al., 2015), confirming that this product addresses real challenges in the field and is ready for widespread adoption.

Table 5. Results From Teacher User Responses

No	Assessment Aspects	Average score	Percentage (%)	Category
1.	Material	4,43	88,50%	Very worthy
2.	Language	4,45	89%	Very worthy
3.	Appearance	4,40	88%	Very worthy
4.	Benefit	4,52	90%	Very worthy
Total score		4,45	89%	Very worthy

These highly positive user reviews are significant. These findings confirm that the product not only meets technical feasibility standards but also delivers high practical value and addresses real-world challenges. Consistent scores above 88% across all aspects, particularly the 90% Benefit aspect, demonstrate that the website effectively assists teachers in grading essays, saving time, and providing more in-depth feedback to students. This success aligns with previous research, such as that by Ramalingan et al. (2018), which also demonstrated that technology can be an effective solution to essay grading challenges. However, this study goes further by validating the product directly with end-users, confirming its acceptability in a real-world setting.

These findings also align with the urgency of modernizing evaluation systems in schools. The response of 89% of teachers who felt a "great need" indicates that there are still significant gaps in current conventional assessment methods. The adoption of machine learning technology for essay assessment can bridge the gap between teachers' need for efficiency and students' need for fast and accurate feedback. Therefore, this research not only provides technical contributions to application development but also provides important insights into how technology can impact educational practices in schools, particularly in physics (Agyei et al., 2024). Overall, this research not only successfully developed a product but also validated its feasibility and need within the educational

community. With strong support from experts and users (teachers), this website shows great potential for replication and further development.

CONCLUSION

This study successfully designed, developed, and validated a machine learning–based web platform for assessing students' essays on Newton's Laws of Motion, employing a Research and Development (R&D) methodology guided by the ADDIE model. The two-stage expert validation demonstrated progressive enhancement in product quality, with Stage 1 yielding an average score of 3.51 (70.2%), categorized as *feasible*, and Stage 2 achieving 3.95 (79.04%), categorized as *highly feasible*. Complementing these results, the field evaluation conducted with assessment-expert teachers produced an average score of 4.45 (89%), indicating a *very high level of practicality and usability*. These findings substantiate that the developed system is pedagogically effective, technically reliable, and well-aligned with classroom assessment demands. The machine learning module not only accelerates the evaluation process but also enhances feedback precision, thereby supporting teachers in delivering more diagnostic, evidence-based insights into student performance. Moreover, the system's integration of automated analytics with qualitative assessment principles exemplifies how artificial intelligence can be harmonized with pedagogical judgment to advance the validity and fairness of educational evaluation.In broader terms, this research contributes to the growing body of knowledge on AI-assisted assessment in science education by demonstrating a replicable framework for technology-enhanced formative evaluation. The developed product holds substantial potential for improving assessment efficiency, promoting reflective learning, and elevating the overall quality of physics education. Future studies are encouraged to extend its implementation across diverse topics and educational contexts to further validate its scalability and pedagogical impact.

REFERENCES

Aizikovitsh-Udi, E., & Cheng, D. (2015). *Developing critical thinking skills from dispositions to abilities: Mathematics education from early childhood to high school*. Creative. http://dx.doi.org/10.4236/ce.2015.64045

Agyei, E., Jita, L., & Jita, T. (2024). Technology integration in science classrooms: Empowering student teachers for improved physics teaching with simulations. *Contemporary Mathematics and Science Education*. https://doi.org/10.30935/conmaths/14688

Akram, H., Abdelrady, A., Al-Adwan, A., & Ramzan, M. (2022). Teachers' Perceptions of Technology Integration in Teaching-Learning Practices: A Systematic Review. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.920317

Al-Mansouri, J. (2024). The Impact of Real-Time Feedback on Optimizing Teachers' Classroom Teaching Pace. *Research and Advances in Education*. https://doi.org/10.56397/rae.2024.11.06

Amanda, F. D., Dewi, U. P., Mufit, F., & Festiyed, F. (2023). The Influence of Essay Assessment on Student Competency Achievement in Science Learning: Literature

Review. *Jurnal Penelitian Pendidikan IPA*, *9*(9), 539-549. https://doi.org/10.29303/jppipa.v9i9.4994

Ariely, M., Nazaretsky, T., & Alexandron, G. (2024). Causal-mechanical explanations in biology: Applying automated assessment for personalized learning in the science classroom. *Journal of Research in Science Teaching*. https://doi.org/10.1002/tea.21929

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning, and Assessment*, *4*(3), 1–30. https://ejournals.bc.edu/index.php/jtla/article/view/1650

Barrow, R. (2015). *Understanding skills: Thinking, feeling, and caring 1st edition*. London: Routledge. https://doi.org/10.4324/9781315678276

Basith, S., Manavalan, B., Shin, T. H., & Lee, G. (2019). SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Molecular Therapy-Nucleic Acids*, *18*, 131-141. https://doi.org/10.1016/j.omtn.2019.08.011

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Blundell, C. (2021). Teacher use of digital technologies for school-based assessment: a scoping review. *Assessment in Education: Principles, Policy & Practice*, 28, 279 - 300. https://doi.org/10.1080/0969594x.2021.1929828

Cahyadi, R. A. H. (2019). Pengembangan bahan ajar berbasis ADDIE model. *Halaqa: Islamic Education Journal*, *3*(1), 35-42. https://doi.org/10.21070/halaqa.v3i1.2124

Carney, M., Webster, B., Alvarado, I., Phillips, K., Howell, N., Griffith, J., ... & Chen, A. (2020). Teachable machine: Approachable Web-based tool for exploring machine learning classification. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems* (pp. 1-8). https://dl.acm.org/doi/abs/10.1145/3334480.3382839

Collins, R. (2014). Skills for the 21st Century: Teaching higher-order thinking. *Curriculum & Leadership Journal*, 12(14), 1-7.

Ding, Y., Zhu, G., Bian, Q., & Bao, L. (2024). Analysis of students' conceptual change in learning Newton's third law with an integrated framework of model analysis and knowledge integration. *Physical Review Physics Education Research*. https://doi.org/10.1103/physrevphyseducres.20.020141

Estrada-Molina, O., Fuentes-Cancell, D., & Morales, A. (2021). The assessment of the usability of digital educational resources: An interdisciplinary analysis from two systematic reviews. *Education and Information Technologies*, 27, 4037 - 4063. https://doi.org/10.1007/s10639-021-10727-5

Gabon, D. (2025). Automated Grading of Essay Using Natural Language Processing: A Comparative Analysis with Human Raters Across Multiple Essay Types. *Journal of Information Systems Engineering and Management*. https://doi.org/10.52783/jisem.v10i6s.700

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education, 57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Gombert, S., Fink, A., Giorgashvili, T., Jivet, I., Di Mitri, D., Yau, J., Frey, A., & Drachsler, H. (2024). From the Automated Assessment of Student Essay Content to Highly Informative Feedback: a Case Study. *Int. J. Artif. Intell. Educ.*, 34, 1378-1416. https://doi.org/10.1007/s40593-023-00387-6

Heil, J., & Ifenthaler, D. (2023). Online Assessment in Higher Education: A Systematic Review. *Online Learning*, *27*(1), 187-218. https://doi.org/10.24059/olj.v27i1.3398

Huang, W., & Fang, N. (2022). Development of a machine learning-based automated assessment system for engineering problem-solving and writing. *International Journal of Engineering Education*, *38*(2), 456–468.

Huang, Q., & Chen, J. (2024). Enhancing academic performance prediction with temporal graph networks for massive open online courses. *Journal of Big Data*, 11, 1-26. https://doi.org/10.1186/s40537-024-00918-5

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208. http://dx.doi.org/10.7717/peerj-cs.208

Indrastoeti, J., & Istiyati, S. (2017). *Asesmen dan evaluasi pembelajaran di sekolah dasar*. UNS Press.

Kamiludin, K., & Suryaman, M. (2017). Problematika pada pelaksanaan penilaian pembelajaran Kurikulum 2013. *Jurnal Prima Edukasia*, *5*(1), 58-67. https://doi.org/10.21831/jpe.v5i1.8391

Kim, N. J., Belland, B. R., & Walker, A. E. (2018). Effectiveness of computer-based scaffolding in the context of problem-based learning for STEM education: Bayesian meta-analysis. *Educational Psychology Review*, *30*(2), 397-429. https://doi.org/10.1007/s10648-017-9419-1

Koe, L., Kustandi, C., & Siregar, E. (2025). AI-driven feedback system: Implementing advanced NLP and openAI for online learning. *Jurnal Inovasi dan Teknologi Pembelajaran*. https://doi.org/10.17977/um031v11i32024p137

Loewenthal, K., & Lewis, C. A. (2018). An introduction to psychological tests and scales. London: Psychology Press.

Loureiro, P., & Gomes, M. J. (2022). The impact of online peer assessment on student learning in higher education: A systematic review of literature. *EDULEARN22 Proceedings*, 4490-4496. https://doi.org/10.21125/edulearn.2022.1074

Luo, Y. (2024). Enhancing educational interfaces: Integrating user-centric design principles for effective and inclusive learning environments. *Applied and Computational Engineering*. https://doi.org/10.54254/2755-2721/64/20241427

Matt C., Charles L. (2015). *Learning Flask Framework Build dynamic, data-driven websitesites and modern website applications with Flask.* Published by Packt Publishing Ltd. Livery Place 35 Livery Street Birmingham B3 2PB, UK.

Marchisio, M., Barana, A., Fioravera, M., Rabellino, S., & Conte, A. (2018). A Model of Formative Automatic Assessment and Interactive Feedback for STEM. *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, 01, 1016-1025. https://doi.org/10.1109/compsac.2018.00178

Mbusi, N., & Luneta, K. (2023). Implementation of an Intervention Program to Enhance Student Teachers' Active Learning in Transformation Geometry. *SAGE Open*, 13. https://doi.org/10.1177/21582440231179440

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, *9*(3), e3292. https://doi.org/10.1002/rev3.3292

Newcombe, N. S., & Shipley, T. F. (2014). *Studying visual and spatial reasoning for design creativity: Thinking about spatial thinking*. Dordrecht: Springer.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Oktaviyanti, I., & Rosyidah, A. N. K. (2019). Korelasi antara Hasil Tes Lisan dengan Hasil Tes Tertulis pada siswa PGSD UNRAM. *Jurnal Ilmu Pendidikan, 2*(1), 9-19. https://doi.org/10.33366/ilg.v2i1.1514

Ole, F. C. B. (2020). Development and Validation of Teachers' Practices on Formative Assessment Scale (TPFAS): A Measure Using Feedback Loop Model. *International Journal of Education,* 13(1), 53–62. https://doi.org/10.17509/ije.v13i1.24715

Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, *32*(2), 604-624. https://doi.org/10.1109/TNNLS.2020.2979670

Perkasa, D. A., Saputra, E., & Fronita, M. (2015). Sistem Ujian Online Essay Dengan Penilaian Menggunakan Metode Latent Sematic Analysis (LSA). *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi*, *1*(1), 1-9. https://doi.org/10.24014/rmsi.v1i1.1313

Plasencia-Calaña, Y. (2025). Operationalizing Automated Essay Scoring: A Human-Aware Approach. *arXiv preprint arXiv:2506.21603*. https://doi.org/10.48550/arXiv.2506.21603

Ramalingam, V., Pandian, A., Chetry, P., & Nigam, H. (2018). Automated Essay Grading using Machine Learning Algorithm. *Journal of Physics: Conference Series,* 1000. https://doi.org/10.1088/1742-6596/1000/1/012030

Ramesh, P., & Sanampudi, S. K. (2023). Interpretable machine learning models for automated essay scoring. *Education and Information Technologies, 28*(5), 5793–5814. https://doi.org/10.1007/s10639-022-11560-1

Retnoningsih E, Pramudita R. (2020). Mengenal Machine Learning Dengan Teknik Supervised dan Unsupervised Learning Menggunakan Python. *Bina Insani ICT Journal*. 7(2): 156-165. http://dx.doi.org/10.51211/biict.v7i2.1422

Ruseti, S., Paraschiv, I., Dascalu, M., & McNamara, D. (2024). Automated Pipeline for Multi-lingual Automated Essay Scoring with ReaderBench. *Int. J. Artif. Intell. Educ.*, 34, 1460-1481. https://doi.org/10.1007/s40593-024-00402-4

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, *35*(5), 535–550. https://doi.org/10.1080/02602930903541015

Sedrakyan, G., Malmberg, J., Verbert, K., Järvelä, S., & Kirschner, P. (2020). Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation. *Comput. Hum. Behav.*, 107, 105512. https://doi.org/10.1016/j.chb.2018.05.004

Shermis, M. D., & Burstein, J. (2019). Handbook of automated essay evaluation: Current applications and new directions (2nd ed.). Routledge. https://doi.org/10.4324/9780429467377

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, *33*(1), 1-19. http://dx.doi.org/10.1111/jcal.12172

Vashishth, T. K., Sharma, V., Sharma, K. K., Kumar, B., Panwar, R., & Chaudhary, S. (2024). AI-driven learning analytics for personalized feedback and assessment in higher education. In Using traditional design methods to enhance AI-driven decision making (pp. 206-230). IGI Global Scientific Publishing. http://dx.doi.org/10.4018/979-8-3693-0639-0.ch009

Vittorini, P., Menini, S., & Tonelli, S. (2020). An AI-Based System for Formative and Summative Assessment in Data Science Courses. I*nternational Journal of Artificial Intelligence in Education*, 31, 159 - 185. https://doi.org/10.1007/s40593-020-00230-2

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, *56*(1), 150-166.https://doi.org/10.1111/bjet.13494

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, *13*(3), 55-75. https://doi.org/10.1109/MCI.2018.2840738

Yu, M., & Tsai, M. (2021). ACS: Construction Data Auto-Correction System—Taiwan Public Construction Data Example. Sustainability. https://doi.org/10.3390/su13010362

U. Vashishth, Sharma, S., & Singh, P. (2024). Artificial intelligence in formative assessment: A review of applications and challenges. *Computers & Education: Artificial Intelligence*, *7*(1), 100212. https://doi.org/10.1016/j.caeai.2023.100212

Zupanc, K., & Bosnić, Z. (2020). Automated essay scoring: A survey of the state of the art. *IEEE Transactions on Learning Technologies*, *13*(4), 821–840. https://doi.org/10.1109/TLT.2020.2996890