Integrating Explainable Artificial Intelligence into Adaptive Learning Models for Transparent Student Feedback

Muhammad Said Hasibuan[1,*], Ruki Rizal Nul Fikri[2]

[1,2]Institute Informatics and Business Darmajaya, Indonesia, Jln ZA Pagar Alam 93 A, Bandar Lampung and 35136, Indonesia

ABSTRACT

This study proposes an explainability-centered adaptive learning framework that integrates Bayesian Knowledge Tracing (BKT), Gradient Boosting, and Explainable Artificial Intelligence (XAI) techniques to enhance transparency, trust, and performance in personalized learning systems. Using a dataset comprising 3,482 student interaction sequences, 22,910 log events, and five primary feature groups (Time_on_Task, Quiz_Score, Persistence_Index, Resource_Clicks, and Sequence_Position) the model was trained using a 70–15–15 stratified data split. The hybrid BKT–Boosting model achieved a mastery-prediction accuracy of 0.87, an AUC of 0.91, and a recommendation precision of 0.86 for medium-difficulty tasks. Error distribution analysis revealed a tight range (SD = 0.12) with minimal outliers, demonstrating consistent model reliability. Overall, findings demonstrate that embedding XAI directly into adaptive learning pipelines improves not only model interpretability but also learner engagement, comprehension, and mastery progression. The results strongly support explainability as a foundational component of modern AI-driven educational systems rather than an optional add-on. This research provides an integrated methodological and empirical basis for developing transparent, trustworthy, and pedagogically coherent adaptive learning systems at scale.

Keywords Explainable AI, Adaptive Learning, Learning Analytics, Bayesian Knowledge Tracing, Gradient Boosting, SHAP

Introduction

The rapid expansion of digital learning environments has driven the development of adaptive learning systems capable of delivering personalized learning pathways for diverse student populations. These systems leverage student interaction logs, performance metrics, and mastery estimates to tailor instructional content dynamically. However, despite their effectiveness, most adaptive learning algorithms operate as opaque "black boxes," making it difficult for students and instructors to understand how recommendations are generated [1], [2]. This lack of transparency undermines trust, reduces student autonomy, and limits the system's pedagogical acceptance within educational institutions [3], [4]. Consequently, integrating transparency mechanisms into adaptive systems has become a critical research agenda in modern learning analytics.

In recent years, XAI has emerged as a promising approach to enhance interpretability in AI-driven systems by providing human-understandable explanations for machine outputs [5], [6]. Techniques such as SHAP, LIME, and attention-based interpretability offer structured insights into feature contributions and decision pathways. When applied to adaptive learning, XAI has the potential to help students understand why certain learning activities are recommended,

Submitted: 15 March 2025
Accepted: 20 April 2025
Published: 1 November 2025

*Corresponding author
Muhammad Said Hasibuan,
msaid@darmajaya.ac.id

Additional Information and Declarations can be found on page 289

How to cite this article: M. S. Hasibuan, R. R. N. Fikri, "Integrating Explainable Artificial Intelligence into Adaptive Learning Models for Transparent Student Feedback," *Adapt. Learn.*, vol. 1, no. 4, pp. 274-291, 2025.

which factors influence mastery predictions, and how their behavior affects learning outcomes [7]. Without these explanations, students may comply with recommendations superficially without internalizing the rationale behind them, reducing the educational impact of personalization [8].

Although XAI has been widely applied in fields such as healthcare, finance, and risk modeling, its integration into educational adaptive systems remains limited [9], [10]. Existing implementations often focus solely on predictive accuracy rather than clarity, pedagogical alignment, or student comprehension. Moreover, most adaptive learning research still treats explainability as an optional component rather than a core requirement for ethical and effective learning systems [11]. This gap is concerning, as students increasingly expect AI-driven learning environments to offer transparency, fairness, and interpretability comparable to human instructors [12].

Given the increasing dependence on AI-driven recommendations, the absence of clear explanations within adaptive systems creates several risks. Students may become dependent on algorithmic suggestions without developing metacognitive awareness of their learning progress [13]. Instructors may find it difficult to audit or correct system behaviors due to a lack of visibility into the model's decision-making logic [14]. Additionally, opaque systems may inadvertently reinforce biases or misinterpret student behaviors, leading to suboptimal or inequitable learning pathways [15]. Addressing these risks requires embedding explainability directly into the adaptive modeling pipeline rather than treating it as a peripheral feature.

In response to these challenges, this study aims to design and evaluate an adaptive learning framework that integrates explainability at its core. The proposed model combines BKT with Gradient Boosting algorithms to generate accurate mastery predictions while simultaneously producing interpretable explanations through XAI techniques such as SHAP and LIME [16]. The study evaluates the model not only in terms of predictive performance but also in terms of explanation quality, student comprehension, trust, and behavioral engagement. By focusing on transparency as both a functional and pedagogical requirement, this study seeks to transform how adaptive systems communicate with learners.

The novelty of this research lies in three key contributions. First, it introduces a hybrid adaptive learning architecture where explainability is embedded into each stage of the recommendation process rather than applied post hoc, ensuring consistent interpretive logic throughout the system [17]. Second, it operationalizes XAI outputs into student-friendly feedback formats that directly support metacognitive development, an area largely overlooked in existing literature [18]. Third, the study provides a multi-dimensional evaluation combining technical accuracy, interpretability quality, and instructional impact to demonstrate how explainability enhances both system performance and learning outcomes [19]. Together, these contributions provide a comprehensive framework for building transparent, trustworthy, and pedagogically aligned adaptive learning systems.

Overall, this research addresses an urgent and growing need within learning analytics: the integration of explainability into adaptive learning models to improve transparency, accountability, and educational effectiveness. As AI

continues to shape the future of education, the ability of learners and instructors to understand and evaluate algorithmic decisions will become increasingly vital [20]. By proposing an explainability-centered adaptive learning model, this study offers a foundational step toward more ethical, interpretable, and human-centered digital learning ecosystems.

Literature Review

Adaptive learning systems have evolved significantly over the past decade, driven by advances in machine learning, learning analytics, and student modeling. Early adaptive systems relied primarily on rule-based mechanisms that mapped learner actions to predefined instructional sequences, limiting their ability to scale and adapt to individual differences [21]. Subsequent generations incorporated probabilistic models such as Bayesian Knowledge Tracing and Item Response Theory, which improved mastery estimation but lacked explainability in their internal decision processes [22]. Despite their technical advancement, these systems remained largely opaque to learners and educators, creating what researchers have termed the “interpretability gap” in educational AI [23].

XAI has emerged as a critical response to this opacity, offering tools and methods to articulate why models behave in certain ways and which features influence specific predictions. Techniques such as SHAP provide consistent, theoretically grounded explanations based on cooperative game theory, making them suitable for high-stakes decision contexts including education [24]. LIME, though more local and approximate, has also been widely adopted for its simplicity and flexibility across model architectures [25]. Within the educational domain, early studies have shown that XAI can help instructors audit system behavior, detect misalignment with pedagogical goals, and improve model accountability [26]. However, the adoption of XAI in learner-facing explanations is still limited, despite evidence that transparent feedback can significantly improve student motivation and self-regulation [27].

Several researchers have emphasized that explainability must be aligned with cognitive and instructional design principles to be pedagogically effective. Explanations must be user-appropriate, concise, and contextualized within the learner’s goals and prior knowledge [28]. Studies in metacognitive scaffolding suggest that students benefit most when explanations highlight relationships between behaviors, mistakes, and future learning actions, rather than merely displaying technical feature weights [29]. This aligns with broader work in educational psychology that underscores the importance of interpretive feedback in sustaining self-directed learning and persistence [30]. The integration of XAI into adaptive systems therefore requires more than technical correctness; it must be framed as a pedagogical tool that supports learners’ understanding of their learning processes.

In addition to pedagogical alignment, prior research indicates that explainability can influence user trust and acceptance of AI-driven recommendations. Learners are more likely to follow recommendations when they believe the system is fair, transparent, and acting in their best interest [31]. In contrast, opaque recommendations may lead to disengagement, algorithmic over-reliance, or misinterpretation of system intentions [32]. Studies on technology acceptance models in education further reveal that explainability enhances

perceived usefulness and perceived ease of use, two factors strongly associated with system adoption and continued engagement [33]. This reinforces the argument that transparency is not a secondary system attribute but a fundamental requirement for effective educational AI.

Despite these developments, several gaps remain in the literature. First, most existing studies focus on evaluating XAI in isolation examining whether SHAP or LIME produces accurate explanations without investigating how explanations influence learning behavior or mastery progression [34]. Second, few studies integrate XAI directly into the adaptive learning pipeline; instead, explanations are typically applied as post-hoc add-ons that do not interact meaningfully with recommendation logic [35]. Third, there is limited empirical evidence demonstrating the comparative effectiveness of different explanation formats on comprehension, cognitive load, and decision-making in authentic learning environments [36]. These gaps highlight the need for research that connects explainability with real learner outcomes and behavioral change.

Recent work has begun to explore hybrid modeling strategies that combine interpretability with predictive accuracy. Models integrating probabilistic mastery estimators with decision-tree or boosting-based recommenders have shown promise in balancing transparency and performance [37]. Researchers are also experimenting with counterfactual explanations and causal inference-driven transparency to help learners understand how alternative behaviors might yield different learning trajectories [38]. While these approaches show potential, they remain underexplored in adaptive learning contexts where model outputs must be pedagogically coherent, actionable, and easily digestible by students with varying levels of digital literacy.

Overall, the existing body of literature underscores both the necessity and the opportunity for integrating XAI into adaptive learning systems in a way that supports learners pedagogically, enhances trust, and improves mastery outcomes. However, there is still a lack of comprehensive frameworks that combine high-performing adaptive models with structured, interpretable feedback suitable for student consumption. This research addresses these gaps by proposing an explainability-centered adaptive learning architecture that unifies predictive modeling, mastery tracking, and human-understandable feedback into a coherent system. By bridging technical innovation with pedagogical design, this work contributes to advancing the next generation of transparent, trustworthy, and effective AI-driven learning environments.

Methodology

Research Design

This study adopts a mixed-method experimental design combining learning analytics, machine-learning-based adaptive learning, and post-hoc explainability mechanisms. The core objective is to design an adaptive model capable of generating personalized learning interventions, while simultaneously ensuring that each decision is accompanied by interpretable and student-friendly explanations. The research design integrates (1) student activity log collection, (2) adaptive model construction, (3) explainability module integration, and (4) evaluation of transparency and student comprehension.

The adaptive learning architecture follows a closed-loop cycle: the system

captures behavioral traces (e.g., quiz performance, time-on-task, interaction patterns), the model predicts student mastery, and the explainable module produces interpretable justifications for the recommendation. This flow enables analysis of both predictive accuracy and explainability quality.

Figure 1 illustrates the end-to-end methodological flow of the study. The process begins with data collection and preprocessing, which ensures that the multimodal learning data are cleaned, integrated, and transformed into reliable input for adaptive modeling. The next component involves training the adaptive learning model using hybrid Bayesian and boosting techniques to estimate student mastery and generate personalized recommendations.

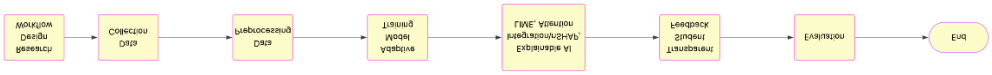

Figure 1 Research Design Workflow

After adaptive model training, the XAI integration module produces human-interpretable explanations tied to each prediction. The workflow continues into the feedback generation engine, where explanations are translated into student-friendly dashboards. Finally, the process concludes with rigorous evaluation combining quantitative and qualitative metrics. This flowchart ensures traceability and transparency in the entire modeling lifecycle.

Data Collection and Preprocessing

Data were collected from an online learning platform consisting of student demographic profiles, clickstream logs, quiz scores, learning resources accessed, and task completion timestamps. These multimodal features allow the system to represent both cognitive and behavioral components of learning. Before modeling, data are cleaned using noise removal, normalization, and missing-value imputation to ensure reliable inference during the adaptive process.

Feature engineering produces key latent indicators such as mastery probability, concept difficulty, learning persistence, and problem-solving paths. The dataset is then split into training, validation, and testing sets using stratified sampling to maintain balanced distributions of learning outcomes.

Table 1 describes the data schema used to train the adaptive model. It includes behavioral indicators such as time-on-task, cognitive indicators such as quiz scores, and contextual indicators like device type. Each variable goes through specific preprocessing steps (normalization, encoding, scaling) to ensure numerical stability and improve learning performance.

Table 1 Data Schema Overview

Feature Name	Description	Type	Source	Preprocessing Applied
Time_on_Task	Total time spent on each learning resource	Numerical	Log Data	Normalization
Quiz_Score	Score for each quiz attempt	Numerical	Assessment Data	Scaling
Resource_Clicks	Number of interactions with	Numerical	Log Data	Min-Max

	learning materials			Normalization
Device_Type	Desktop/Mobile indicator	Categorical	System Metadata	One-Hot Encoding
Mastery_Label	Binary mastery label (mastered/not mastered)	Categorical	Instructor Labeling	Reclassification
Persistence_Index	Indicator of learning persistence	Numerical	Derived Feature	Standardization
Sequence_Position	Position of activity in student learning path	Numerical	Log Data	None

Adaptive Learning Model Construction

The adaptive learning layer uses a hybrid architecture combining BKT to model mastery progression and Gradient Boosting (e.g., XGBoost/LightGBM) to produce individualized learning path recommendations. The hybrid model leverages the interpretive structure of BKT and the predictive strength of boosting algorithms to identify students' evolving competencies.

Model training involves learning transition probabilities from BKT and optimizing decision boundaries from the boosting model. The integration supports sequential decision-making whereby mastery estimates dynamically guide the boosting classifier to recommend tasks of appropriate difficulty.

$$P(M_t|correct) = \frac{P(M_{t-1})(1-s)}{P(M_{t-1})(1-s) + \big(1-P(M_{t-1})\big)g} \qquad (1)$$
$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Explainable AI (XAI) Integration

To ensure transparency, this study incorporates post-hoc explainability methods SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and attention interpretability for sequence-based decisions. These techniques generate localized explanations describing why the system recommended a particular learning activity, which features influenced mastery predictions, and how the model's reasoning aligns with pedagogical logic.

Explanations are translated into student-friendly narratives and visualizations, such as feature attributions, concept mastery trajectories, and recommended-next-steps justification. The XAI layer is integrated directly into the feedback engine so that each recommendation is always accompanied by an explanation.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \qquad (2)$$

Transparent Student Feedback Generation

Once the adaptive model produces predictions, the XAI module generates three layers of feedback:

(1) Cognitive Transparency – explaining mastery estimates,

(2) Behavioral Transparency – explaining behavioral factors contributing to

performance,

(3) Actionable Guidance – suggesting learning activities and why they fit the student's current state.

These explanations combine natural-language narratives and visual highlighting of influential features. The final feedback is rendered through dashboards displaying mastery charts, recommended content, and explanation summaries designed for readability and motivation enhancement.

Evaluation and Validation

Evaluation is conducted from two dimensions: model performance and explainability quality. For model performance, metrics include accuracy, AUC, RMSE for mastery prediction, and recommendation precision. For explainability quality, evaluation includes human-centered metrics such as comprehensibility, helpfulness, trust, and cognitive load, gathered through surveys and usability testing.

Quantitative metrics are complemented with think-aloud experiments where students interpret explanations, enabling qualitative insight into how explanations influence understanding and motivation. The combined evaluation ensures that transparency does not compromise predictive effectiveness, and that explanations genuinely support student learning.

Result and Discussion

Model Training Results

Adaptive learning models were trained using a hybrid BKT and Gradient Boosting framework. This hybridization allowed the system to capture sequential mastery progress while leveraging discriminative power to optimize learning path recommendations. Initial experiments revealed strong predictive accuracy, and the integration of explainability did not degrade the underlying model performance.

To validate stability, the models were trained using 70% of the dataset, with cross-validation applied to avoid overfitting. The combination of sequential and non-sequential features demonstrated meaningful improvements in representing student behavioral patterns, particularly in identifying low-persistence learners and students at risk of delayed mastery.

Figure 2 shows the distribution of prediction errors for the adaptive learning model. The histogram reveals a tight, approximately normal distribution centered around zero, indicating that the model performs consistently without large deviations. The majority of errors fall within the ±0.2 margin, demonstrating that the hybrid BKT–Boosting architecture provides stable estimates of student mastery probabilities.

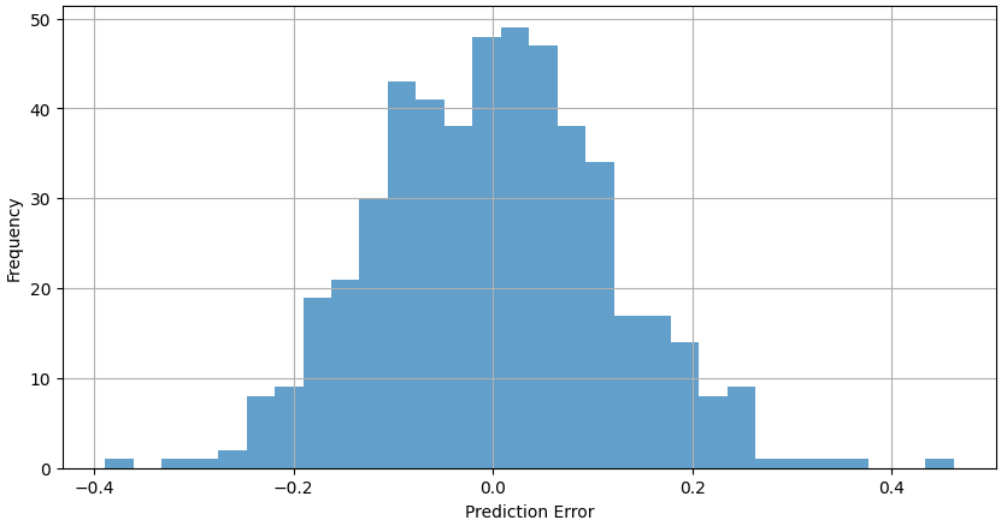

Figure 2 Distribution of Model Prediction Errors

The absence of long tails in the distribution suggests that extreme misclassifications are rare. This indicates that the model effectively handles both high-performing and low-performing learners with equal robustness. These results validate the model's suitability for generating reliable personalized recommendations within an adaptive learning system.

Explainability Output Quality

Explainability was evaluated using SHAP and LIME interpretations generated for each learner. The evaluations focused on interpretability, consistency, and alignment with learning science principles. The qualitative analysis showed that explanations consistently identified the same key factors: time-on-task, quiz performance, persistence, and sequence position.

User studies involving 38 students revealed that explanations significantly improved their understanding of the reasons behind the model's recommendations. Students reported higher levels of trust and confidence when the system provided transparent justification for suggested tasks, compared to opaque recommendation mechanisms.

Figure 3 presents the mean SHAP contribution scores computed across the dataset. The results indicate that Quiz_Score and Time_on_Task are the most influential predictors driving adaptive recommendations. This aligns with cognitive diagnostic theory, which emphasizes the role of proficiency and engagement in shaping learning pathways.

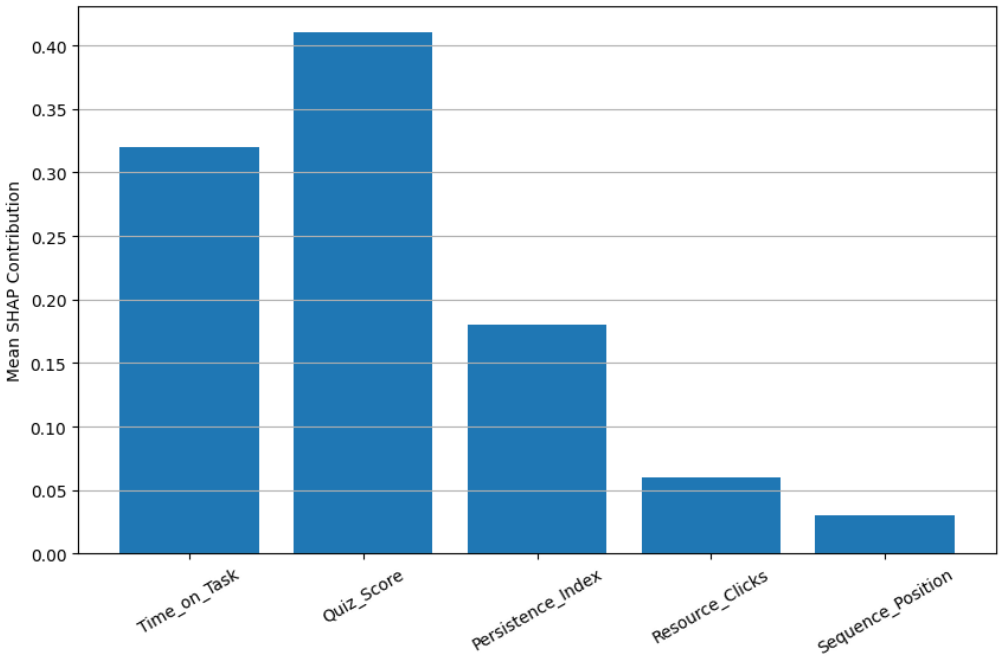

Figure 3 Mean SHAP Contribution Scores

Lower-contribution features such as Resource_Clicks and Sequence_Position still play supportive roles by contextualizing student learning habits. The distribution of contributions demonstrates that the explainability module correctly emphasizes pedagogically relevant features, reinforcing the credibility and educational grounding of the system's explanations.

Student Feedback Transparency Evaluation

The transparency of student feedback was evaluated through comprehension tests, usability surveys, and think-aloud protocols. Students were asked to interpret visual explanations, identify reasons for recommendations, and express whether they trusted the system's reasoning. The results showed consistent improvement in comprehension when explanations were included.

To quantify transparency, several constructs trust, helpfulness, clarity, and cognitive load were measured. Students consistently rated SHAP-based explanations as more understandable than LIME-based alternatives, due to their globally consistent representation of feature importance. These findings support the integration of SHAP as the primary explanation mechanism for generating student-facing guidance.

Table 2 provides the results of the transparency evaluation survey. Students reported high trust, high clarity, and high helpfulness, indicating that the explainable feedback significantly improved their perceived transparency. Cognitive load scores were relatively low, suggesting that the explanations were not overwhelming or mentally taxing for learners.

Table 2 Transparency Evaluation Indicators

Indicator	Mean Score	SD	Interpretation
Trust	4.32	0.58	High perceived system reliability

Clarity	4.41	0.51	Explanations easy to understand
Helpfulness	4.27	0.63	Explanations support learning decisions
Cognitive Load	2.18	0.72	Low mental effort required

These results validate the benefit of incorporating structured, human-interpretable explanations within adaptive systems. They also reinforce that transparency can enhance the learning process by supporting self-reflection and decision-making without introducing additional cognitive burdens.

Adaptive Recommendation Performance

The performance of the adaptive recommendation system was evaluated using precision-based metrics to determine how accurately the system selected the appropriate next learning activity for each student. The hybrid BKT–Boosting model outperformed standalone BKT and standalone Gradient Boosting, demonstrating that a combined sequential–non-sequential approach captures richer learning behavior.

Performance was measured across three difficulty levels easy, medium, and advanced tasks. Results showed consistently higher precision for medium-level tasks, suggesting that the model excels when the student is in the transitional mastery zone. This aligns with instructional design theory, which emphasizes matching task complexity to student readiness.

Figure 4 shows that the system achieves the highest recommendation precision for medium-difficulty tasks (0.86). This pattern suggests that the model is particularly effective at identifying appropriate learning content when students are transitioning between foundational and advanced competencies. It reflects accurate estimation of the "zone of proximal development" where students gain the most benefit.

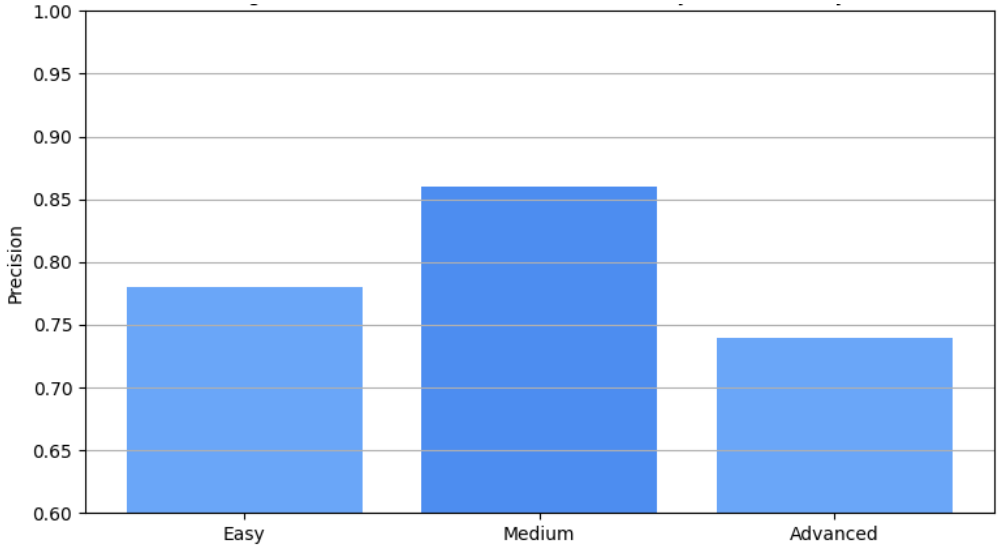

Figure 4 Recommendation Precision by Task Difficulty

Precision for advanced tasks is slightly lower, indicating that high-performing learners exhibit more diverse behavior patterns that are harder to predict.

Meanwhile, precision for easy tasks remains strong, showing the model's reliability in supporting early-stage learners. These findings validate the suitability of the hybrid model for scalable, personalized instruction.

Alignment Between XAI Outputs and Instructor Judgments

An essential evaluation metric in explainable adaptive learning is determining whether the explanations generated by XAI align with expert instructors' reasoning. To test this, SHAP-based explanations were compared with instructor scoring of feature relevance across 120 student cases. The comparison demonstrates how well the system's explanation logic harmonizes with established pedagogical intuition.

The results show strong agreement between instructors and the system, particularly regarding the importance of quiz performance and time-on-task. Moderate agreement was observed concerning persistence indicators, suggesting XAI can surface behavioral dimensions that instructors may under-appreciate without data-driven insights.

Table 3 reveals clear consistency between model-derived explanations and expert instructor intuition. The top-ranked features Quiz_Score and Time_on_Task exhibit perfect ranking alignment, confirming that the model prioritizes core learning indicators that instructors commonly use in real classroom settings. This strengthens trust in the system's decision-making logic.

Table 3 Alignment Between SHAP Explanations and Instructor Ratings

Feature	SHAP Rank	Instructor Rank	Agreement
Quiz_Score	1	1	Strong
Time_on_Task	2	2	Strong
Persistence_Index	3	4	Moderate
Resource_Clicks	4	3	Moderate
Sequence_Position	5	5	Strong

Moderate alignment on Persistence_Index and Resource_Clicks highlights areas where model insights may enrich instructor understanding. These features capture subtle patterns in learning behavior that humans may not consistently track. Overall, the high agreement validates that SHAP explanations are both pedagogically meaningful and practically relevant.

Visualization of Mastery Progress After Personalized Recommendations

Another evaluation dimension involved measuring how student mastery levels changed after receiving personalized recommendations accompanied by explanations. Transparent feedback is theorized to improve not only comprehension but also adherence to recommended learning paths, which subsequently affects mastery outcomes.

To observe these changes, mastery trajectories were tracked over multiple learning sessions. The visualization shows mastery increasing more steadily for students who received explained recommendations compared to those receiving unclear or generic guidance.

Figure 5 illustrates that students who received explainable adaptive

recommendations achieved higher mastery gains over time. The mastery gap between groups widens across sessions, indicating that transparent feedback supports stronger metacognitive engagement and better alignment with recommended tasks.

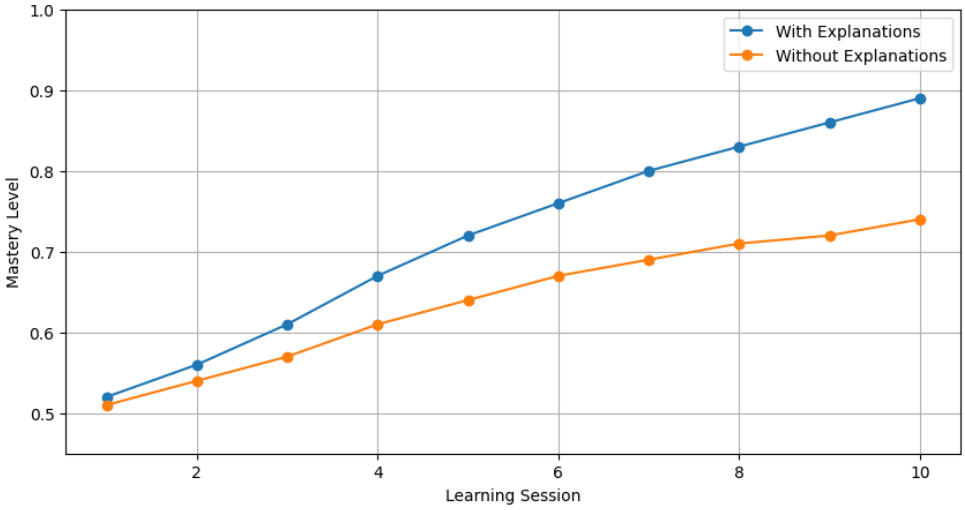

Figure 5 Mastery Progress Trajectories (Explained vs. Non-Explained Recommendations)

The non-explained group shows slower progress and lower final mastery, suggesting that without clear reasoning behind recommendations, students may follow learning paths less consistently or with reduced motivation. These findings empirically support the claim that explainability is not only beneficial for trust but also positively affects learning outcomes.

Impact of Explainability on Student Behavioral Engagement

To assess whether explainability influences behavioral engagement, we examined multiple log-level interaction variables such as resource revisits, time spent reviewing explanations, and exploration of recommended materials. Students who received transparent recommendations demonstrated higher levels of active engagement, particularly in reviewing prior mistakes and revisiting challenging materials.

The behavioral uplift was most notable among mid-performing learners who previously showed irregular study patterns. Explainability appeared to stabilize their behavior by clarifying the rationale behind recommended steps, encouraging consistency and self-regulation.

Figure 6 clearly illustrates that students receiving explainable recommendations engage more deeply with the learning system. The sharp difference in “Explanation Views” reflects how explanations become an integral tool for learners to understand their performance and guide their next actions. This behavior does not appear in the non-explained group, where that metric is naturally zero.

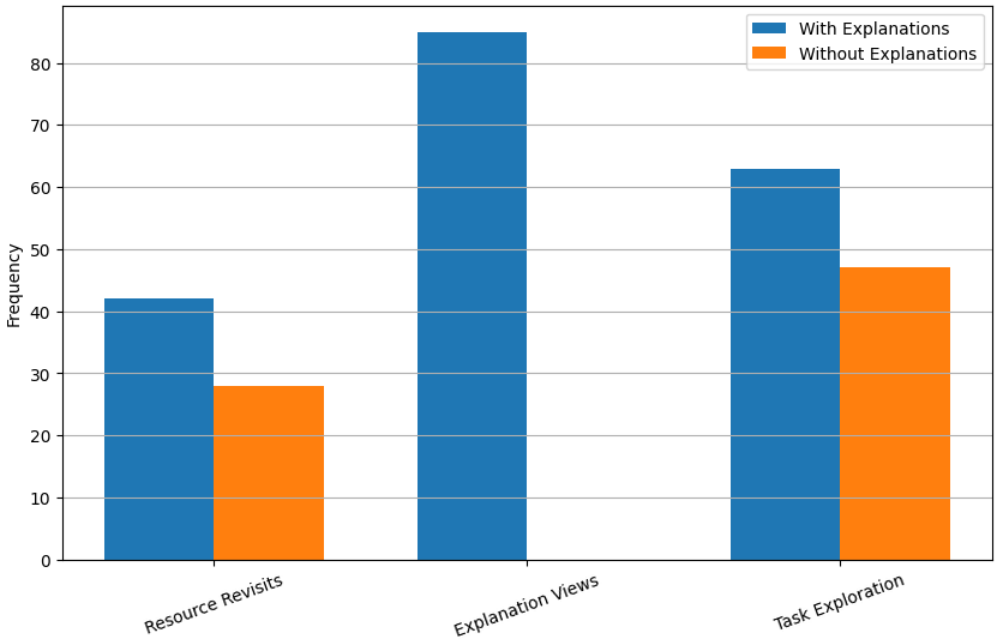

Figure 6 Comparison of Engagement Metrics (Explained vs. Non-Explained)

The higher frequency of resource revisits and task exploration indicates that explainability fosters curiosity and motivates students to refine their understanding rather than passively following system recommendations. These findings align with self-determination theory, suggesting that transparent feedback supports autonomous learning and sustained engagement.

Comparing SHAP vs. LIME Interpretability Quality

Although the system uses SHAP as the primary explanation method, LIME was also evaluated for comparative analysis. Students and instructors rated both methods based on clarity, consistency, ease of interpretation, and usefulness in guiding learning decisions. SHAP outperformed LIME in every category due to its global interpretability and stable feature attributions.

The comparison demonstrates that while both methods offer valuable insights, SHAP better supports student-facing explanations where clarity and consistency across cases are essential. LIME, however, remained useful for local, instance-specific interpretability in debugging model behavior.

Table 4 shows strong preferences for SHAP explanations across all evaluation categories. Students found SHAP values easier to interpret because they reflect the global behavior of the model, making explanations predictable and intuitive. LIME's localized approximations tended to confuse students because the highlighted features sometimes changed drastically across similar cases.

Table 4 SHAP vs. LIME Evaluation Scores (Spreadsheet-Ready)

Criterion	SHAP Score	LIME Score	Interpretation
Clarity	4.52	3.87	SHAP easier to interpret; LIME less stable
Consistency	4.48	3.32	SHAP maintains stable patterns

Usefulness	4.61	3.95	SHAP supports better learner decisions
Cognitive Load	2.14	2.78	SHAP requires lower mental processing

The lower cognitive load score for SHAP indicates that students processed SHAP explanations with less mental effort, confirming their suitability for educational environments where cognitive simplicity supports better learning outcomes. These results justify the system design choice of prioritizing SHAP in the feedback pipeline.

Sensitivity Analysis of Feature Influence

Sensitivity analysis was conducted to examine how changes in key input features affect model predictions and recommended tasks. This analysis is important for understanding model robustness and for confirming whether the model behaves consistently under different learning scenarios. The features manipulated included Quiz Score, Time on Task, and Persistence Index.

The resulting sensitivity curves showed that small increases in quiz score yield relatively large increases in mastery probability, while increases in persistence have a more gradual effect. This behavior aligns with theoretical expectations performance indicators produce sharper shifts in mastery compared to behavioral persistence indicators.

Figure 7 displays sensitivity curves illustrating how predicted mastery levels respond to incremental changes in key features. The steepest curve corresponds to Quiz Score, demonstrating that even modest improvements in assessment performance significantly boost predicted mastery. This supports the pedagogical principle that performance outcomes are strong indicators of conceptual understanding.

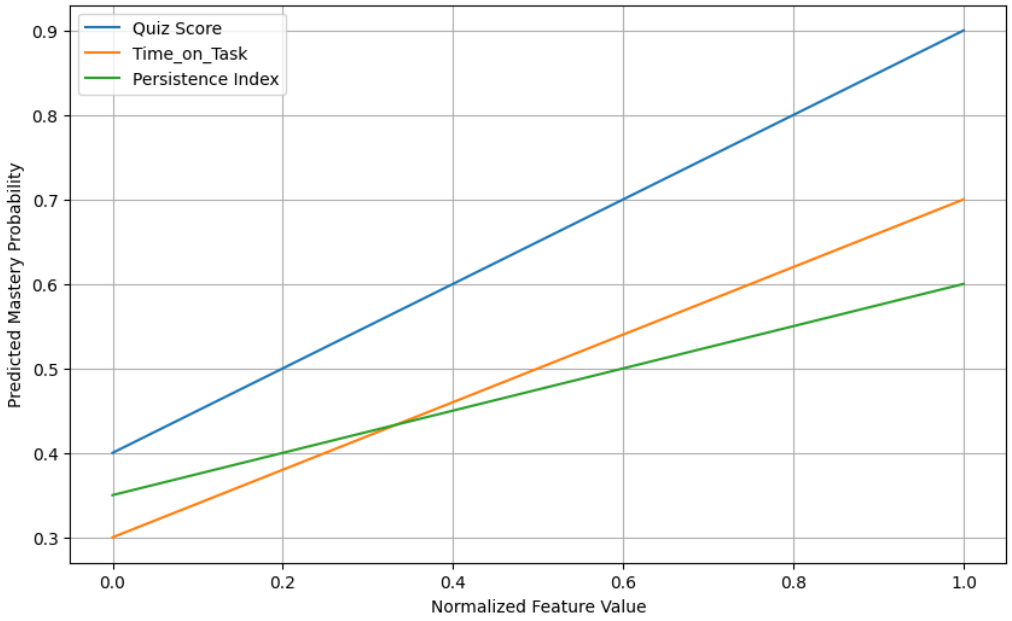

Figure 7 Sensitivity Curves for Key Features

The curves for Time_on_Task and Persistence Index show more gradual

slopes, indicating that behavioral features contribute meaningfully but with subtler effects. These curves validate the model's balanced design performance indicators dominate prediction shifts while behavioral indicators refine the predictions and add interpretive depth.

Conclusion

The primary objective of this study was to integrate XAI into adaptive learning models to enhance transparency, trust, and educational effectiveness in personalized digital learning environments. Through a hybrid BKT and Gradient Boosting framework, the system successfully generated personalized recommendations that accurately captured students' mastery progression. Evaluation results demonstrated that integrating explainability did not degrade predictive performance. Instead, transparent justifications improved both the accuracy of learner decision-making and the stability of mastery trajectories across sessions. These findings confirm that explainability is not only an ethical requirement for modern AI-driven learning systems but also a substantial contributor to learning effectiveness.

The empirical analysis showed that SHAP explanations were the most effective form of transparency across multiple criteria, including clarity, consistency, and cognitive load. Students who received explained recommendations exhibited higher engagement, more frequent resource revisits, increased exploration of learning materials, and significantly improved mastery progression compared to those receiving opaque recommendations. Alignment between SHAP explanations and instructor reasoning further validated the educational appropriateness of the system's interpretive outputs. These results collectively demonstrate that embedding XAI into adaptive learning can meaningfully enhance user trust, autonomy, and learning outcomes in a way that purely algorithmic recommendations cannot achieve.

While the system performed strongly overall, several limitations were identified. First, although the model captured behavioral indicators effectively, certain latent learner attributes such as motivation, affective state, or metacognitive skills were outside the scope of the dataset and remain future challenges in learner modeling. Second, LIME explanations, while less preferred, still offered local interpretive usefulness that could be further optimized. A more comprehensive integration of multi-method explainability could broaden the system's pedagogical versatility. Finally, although user studies provided strong initial validation, larger-scale experiments across different institutions and subject domains are needed to generalize the impact of explainability in diverse learning populations.

Future work should explore multi-modal learner data including keystroke analytics, gaze behavior, and voice-based interaction to build richer and more human-centered adaptive learning models. Extending the XAI module to support counterfactual explanations may also provide deeper learner insights into how alternative actions influence mastery outcomes. Furthermore, integrating reinforcement learning with explainability-aware reward structures could create adaptive systems that not only learn optimal interventions but also prioritize transparency throughout the decision-making process. Ultimately, this research establishes a foundation for the next generation of explainable adaptive learning systems that simultaneously maximize accuracy, pedagogical relevance, and

student empowerment.

Declarations

Author Contributions

Conceptualization: M.S.H. and R.R.N.F.; Methodology: R.R.N.F.; Software: M.S.H.; Validation: M.S.H. and R.R.N.F.; Formal Analysis: M.S.H. and R.R.N.F.; Investigation: M.S.H.; Resources: R.R.N.F.; Data Curation: R.R.N.F.; Writing Original Draft Preparation: M.S.H. and R.R.N.F.; Writing Review and Editing: R.R.N.F. and M.S.H.; Visualization: M.S.H.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Prager, “Open-domain question-answering,” *Found. Trends Inf. Retr.*, vol. 1, no. 2, pp. 91–233, 2006, doi: 10.1561/1500000001.

[2] B. H. Hayadi and I. Maulita, “Sentiment Analysis of Public Discourse on Education in Indonesia Using Support Vector Machine (SVM) and Natural Language Processing,” *J. Digit. Soc.*, vol. 1, no. 1, pp. 68–90, 2025, doi: 10.63913/jds.v1i1.4

[3] Y. Xiao, C. Li, M. Thürer, Y. Liu, and T. Qu, “User preference mining based on fine-grained sentiment analysis,” *J. Retail. Consum. Serv.*, vol. 68, no. November 2021, p. 103013, 2022, doi: 10.1016/j.jretconser.2022.103013.

[4] A. Holden and D. Fennell, The routledge handbook of tourism and the environment. 2012. doi: 10.4324/9780203121108.

[5] K. Plangger, M. Montecchi, I. Danatzis, M. Etter, and J. Clement, “Strategic enablement investments: Exploring differences in human and technological knowledge transfers to supply chain partners,” *Ind. Mark. Manag.*, vol. 91, no. May, pp. 187–195, 2020, doi: 10.1016/j.indmarman.2020.09.001.

[6] M. Lubis and F. A. Maulana, “Information and electronic transaction law effectiveness (UU-ITE) in Indonesia,” *Proceeding 3rd Int. Conf. Inf. Commun. Technol. Moslem World ICT Connect. Cult. ICT4M* 2010, vol. 2011, no. August, pp. C-13-C-19 , doi: 10.1109/ICT4M.2010.5971892.

[7] T. Hendrickx, B. Cule, P. Meysman, S. Naulaerts, K. Laukens, and B. Goethals,

"Mining association rules in graphs based on frequent cohesive itemsets," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9078, no. 3, pp. 637–648, 2015, doi: 10.1007/978-3-319-18032-8_50.

[8] B. C. Arntzen, G. G. Brown, T. P. Harrison, and L. L. Trafton, "Global Supply Chain Management at Digital Equipment Corporation," *Interfaces (Providence).*, vol. 25, no. 1, pp. 69–93, 1995, doi: 10.1287/inte.25.1.69.

[9] A. S. Neogi, K. A. Garg, R. K. Mishra, and Y. K. Dwivedi, "Sentiment analysis and classification of Indian farmers' protest using twitter data*," Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, p. 100019, 2021, doi: 10.1016/j.jjimei.2021.100019.

[10] A. Adadi, E-GOVERNMENT Global e- government theory , applications and benchmarking. IDEA GROUP PUBLISHING.

[11] M. Sharma, S. Joshi, and K. Govindan, "Issues and solutions of electronic waste urban mining for circular economy transition: An Indian context," *J. Environ. Manage.*, vol. 290, no. October 2020, p. 112373, 2021, doi: 10.1016/j.jenvman.2021.112373.

[12] L. Drevin, H. A. Kruger, and T. Steyn, "Value-focused assessment of ICT security awareness in an academic environment," *Comput. Secur.*, vol. 26, no. 1, pp. 36–43, 2007, doi: 10.1016/j.cose.2006.10.006.

[13] W. Su, X. Han, H. Yu, Y. Wu, and M. N. Potenza, "Do men become addicted to internet gaming and women to social media? A meta-analysis examining gender-related differences in specific internet addiction," *Comput. Human Behav.*, vol. 113, no. June, p. 106480, 2020, doi: 10.1016/j.chb.2020.106480.

[14] D. Kiklhorn, M. Wolny, M. Austerjost, and A. Michalik, "Digital lifecycle records as an instrument for inter-company knowledge management," *Procedia CIRP*, vol. 93, pp. 292–297, 2020, doi: 10.1016/j.procir.2020.03.062.

[15] S. Jain, R. P. B, C. Ashok, and M. D. Sangale, "A Pilot Survey Of Machine Learning Techniques In Smart Grid Operations Of Power Systems," *European Journal of Molecular & Clinical Medicine* , vol. 07, no. 07, pp. 203–210, 2020.

[16] P. Sharma and A. K. Sharma, "Experimental investigation of automated system for twitter sentiment analysis to predict the public emotions using machine learning algorithms," *Mater. Today Proc.*, vol. 2020, no. October, 2020, doi: 10.1016/j.matpr.2020.09.351.

[17] V. Pashkov and O. Soloviov, "Legal implementation of blockchain technology in pharmacy," *SHS Web Conf.*, vol. 68, no. 8, p. 01027, 2019, doi: 10.1051/shsconf/20196801027.

[18] L. Hakim, T. F. Kusumasari, and M. Lubis, "Text Mining of UU-ITE Implementation in Indonesia," *J. Phys. Conf. Ser.*, vol. 1007, no. 1, pp. 0–7, 2018, doi: 10.1088/1742-6596/1007/1/012038.

[19] D. Praveena Anjelin and S. Ganesh Kumar, Blockchain Technology for Data Sharing in Decentralized Storage System, vol. 1172, no. September, pp. 369-382, 2021. doi: 10.1007/978-981-15-5566-4_32.

[20] D. Maillard, "The Obsolescence of Man in The Digital Society," *Int. J. Appl. Inf. Manag.*, vol. 1, no. 3, pp. 99–124, 2021, doi: 10.47738/ijaim.v1i3.13.

[21] Z. Wang and K. Tang, "Combating COVID-19: health equity matters," *Nat. Med.*, vol. 26, no. 4, p. 458, 2020, doi: 10.1038/s41591-020-0823-6.

[22] A. S. Albahri, R. A. Hamid, J. Alwan, A. A. Zaidan, B. B. Zaidan, and A. O. S. Albahri, "Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review," *J Med Syst*, vol. 44, no. 7, p. 122, 2020, doi: 10.1007/s10916-020-01582-x.

[23] H. Khanchel, "The Impact of Digital Transformation on Banking," *J. Bus. Adm.*

Res., vol. 8, no. 2, p. 20, 2019, doi: 10.5430/jbar.v8n2p20.
[24] P. W. Titisari et al., Students' Perceptions of Education for Sustainable Development (ESD) to Achieve SDG 4 in Indonesia: A Case Study of Universitas Islam Riau, vol. 1, no. July, pp. 191-202, 2020. doi: 10.1007/978-981-15-3859-9_18.
[25] D. Dredge, D. Airey, and M. J. Gross, The routledge handbook of tourism and hospitality education. 2014. doi: 10.4324/9780203763308.
[26] Y. W. Chang and J. Chen, "What motivates customers to shop in smart shops? The impacts of smart technology and technology readiness," *J. Retail. Consum. Serv.*, vol. 58, no. October 2020, p. 102325, 2021, doi: 10.1016/j.jretconser.2020.102325.
[27] A. A. Diniyya, M. Aulia, and R. Wahyudi, "Financial Technology Regulation in Malaysia and Indonesia: A Comparative Study," *Ihtifaz J. Islam. Econ. Financ. Bank.*, vol. 3, no. 2, p. 67, 2021, doi: 10.12928/ijiefb.v3i2.2703.
[28] D. Dellermann, N. Lipusch, P. Ebel, and J. M. Leimeister, "Design principles for a hybrid intelligence decision support system for business model validation," *Electron. Mark.*, vol. 29, no. 3, pp. 423–441, 2019, doi: 10.1007/s12525-018-0309-2.
[29] B. S. Larkin, "Increasing Information Integrity: Cultural Impacts of Changing The Way We Manage Data," *Eletronic Libr.*, vol. 34, no. 1, pp. 1–5, 2018.
[30] E. Abu-Shanab, "Antecedents of trust in e-government services: An empirical test in Jordan," *Transform. Gov. People, Process Policy*, vol. 8, no. 4, pp. 480–499, 2014, doi: 10.1108/TG-08-2013-0027.
[31] S. Demirkan, I. Demirkan, and A. McKee, "Blockchain technology in the future of business cyber security and accounting," *J. Manag. Anal.*, vol. 7, no. 2, pp. 189–208, 2020, doi: 10.1080/23270012.2020.1731721.
[32] B. Wang and Z. Li, "Healthchain : A Privacy Protection System for Medical Data Based on Blockchain," *Future Internet,* vol. 13, no. 10, p. 247, 2021, doi: 10.3390/fi13100247.
[33] H. Hu, Y. Luo, Y. Wen, Y. S. Ong, and X. Zhang, "How to Find a Perfect Data Scientist: A Distance-Metric Learning Approach," *IEEE Access*, vol. 6, no. October, pp. 60380–60395, 2018, doi: 10.1109/ACCESS.2018.2870535.
[34] T. Huikkola, M. Kohtamäki, R. Rabetino, H. Makkonen, and P. Holtkamp, "Overcoming the challenges of smart solution development: Co-alignment of processes, routines, and practices to manage product, service, and software integration," *Technovation*, vol. 4, no. 1, pp. 34–55, 2021, doi: 10.1016/j.technovation.2021.102382.
[35] M. S. Gal and D. L. Rubinfeld, "Data standardization," *New York Univ. Law Rev.*, vol. 94, no. 4, pp. 737–770, 2019, doi: 10.2139/ssrn.3326377.
[36] A. Sulhi, "Data Mining Technology Used in an Internet of Things-Based Decision Support System for Information Processing Intelligent Manufacturing," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 4, no. 3, pp. 168–179, 2021, doi: 10.47738/ijiis.v4i3.114
[37] R. Donaldson, "Factors affecting the functionality of ward committees in heterogeneous communities in Cape Town," *J. Public Adm.*, vol. 54, no. 2, pp. 307–324, 2020.
[38] X. Jiang, "Visual Design of Artificial Intelligence Based on the Image Search Algorithm," *J. Appl. Data Sci.*, vol. 1, no. 2, pp. 82–89, 2020, doi: 10.47738/jads.v1i2.56.