International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Indexed, Double Blind, Open Access, Peer-Reviewed, Refereed-International Journal.
SJIF Impact Factor =7.938, July-December 2024, Submitted in July 2024, ISSN -2393-8048

Explainable AI for Engagement Prediction in Online and Offline Educational Environments

Sunil Kumar, Research Scholar, Department of Technology and Computer Science, Glocal University, Saharanpur (Uttar Pradesh)
Dr. Amit Singla, Assistant Professor, Department of Technology and Computer Science, Glocal University, Saharanpur (Uttar Pradesh)

Abstract

The integration of Artificial Intelligence (AI) in educational environments has revolutionized the way academic engagement is analyzed and predicted. However, the "black-box" nature of many AI models limits their interpretability, raising concerns about trust, fairness, and accountability. This research proposes an Explainable AI (XAI) framework for predicting student engagement in online and offline learning environments. By leveraging state-of-the-art deep learning models and incorporating explainability techniques such as SHAP (Shapley Additive Explanations), the study aims to enhance the transparency and usability of engagement prediction systems. The paper evaluates the proposed framework through extensive experimentation and a case study in a hybrid learning setup.

Keywords: Artificial Intelligence, Explainable AI, SHAP.

1. Introduction

1.1 Background and Motivation

Explainable Artificial Intelligence (XAI) has become an essential focus in modern AI research, addressing the need for transparency, interpretability, and trustworthiness in AI systems. In the context of educational environments, particularly in engagement prediction, XAI offers the opportunity to unveil the decision-making processes behind AI-driven models, ensuring that educators, students, and stakeholders can understand and trust the outputs. Engagement prediction involves assessing how actively a learner participates in educational activities—whether measured through behavioral, cognitive, or emotional involvement—and is a critical determinant of learning outcomes. This concept bridges psychology, pedagogy, and technology, and accurate predictions can lead to improved instructional strategies and personalized interventions. The rise of online education platforms and the continued importance of traditional offline classrooms have highlighted the growing challenges in maintaining and assessing student engagement. In online environments, where students often learn remotely and asynchronously, traditional cues like body language or real-time feedback from teachers are either absent or limited. Engagement metrics in such settings often rely on data sources such as interaction logs (e.g., clicks, video watch time, or quiz participation), eye-tracking, and sentiment analysis from facial expressions. On the other hand, offline classrooms rely heavily on real-time observation by teachers, which is inherently subjective and prone to bias. In both settings, AI models have been increasingly deployed to predict engagement levels. These models analyze large volumes of data to identify patterns that signify high or low engagement. However, the “black-box” nature of AI—where models provide predictions without clear explanations—poses challenges to trust, usability, and fairness in educational decision-making. This is where XAI plays a transformative role. By integrating explainability into engagement prediction systems, educators can receive insights not only into what the predictions are but also why the model arrived at those predictions. For instance, an XAI-enabled model could reveal that a student’s disengagement is due to the complexity of the material or a lack of interactive elements in a lesson. Such insights empower educators to design targeted interventions, such as simplifying content, adding interactive elements, or providing one-on-one support. Moreover, XAI ensures that biases in AI systems are detected and mitigated, fostering fairness in diverse educational contexts. For example, if an engagement model consistently predicts lower engagement for students from specific socio-economic backgrounds, XAI tools can highlight and address such biases.

The motivation for this study arises from the increasing demand for personalized learning experiences, which adapt to the unique needs, abilities, and preferences of individual learners.

Traditional methods of engagement assessment—such as manual observation, questionnaires, or surveys—are limited in scalability and objectivity. They are often reactive, identifying engagement issues only after they have affected learning outcomes. AI-based systems, on the other hand, can process multimodal data—such as video, audio, text, and interaction logs—in real-time, enabling proactive interventions. However, the lack of transparency in these systems has hindered their widespread adoption, as stakeholders are wary of decisions that lack clarity or justification. Furthermore, the integration of XAI into engagement prediction aligns with the broader educational goals of equity and inclusivity. Transparent and explainable models can highlight systemic issues, such as teaching methods that disproportionately disengage certain groups of students, allowing institutions to address these problems effectively. Additionally, XAI can foster greater student agency by providing learners with insights into their own engagement patterns, empowering them to take an active role in their education.

Importance of Academic Engagement in Educational Outcomes

Academic engagement plays a pivotal role in shaping educational outcomes, acting as a crucial determinant of students' academic achievement, retention rates, and overall development. It encompasses behavioral, emotional, and cognitive dimensions that collectively influence how students interact with learning materials, educators, and peers. Behavioral engagement refers to participation in academic tasks such as attending classes, completing assignments, and contributing to discussions, which are strong predictors of academic success (Fredricks et al., 2004). Emotional engagement, including feelings of belonging, interest, and enthusiasm, fosters intrinsic motivation and resilience, enabling students to persist through academic challenges (Appleton et al., 2008). Cognitive engagement, characterized by deep learning strategies and self-regulated learning, equips students with the ability to think critically and solve complex problems, essential skills in today's knowledge-driven economy (Connell & Wellborn, 1991). Studies indicate that higher engagement levels correlate with improved academic performance and reduced dropout rates, highlighting its integral role in education (Wang & Eccles, 2013). Therefore, fostering an environment that promotes active engagement is a priority for educational institutions seeking to enhance both individual and institutional outcomes.

The integration of Artificial Intelligence (AI) in education is transforming how engagement is assessed, predicted, and enhanced. Advanced AI systems utilize large datasets generated from Learning Management Systems (LMS), video lectures, online discussion forums, and even biometric sensors to analyze and predict engagement patterns in real-time (Zhang et al., 2021). Techniques such as machine learning and deep learning enable the detection of nuanced engagement indicators, including attendance patterns, interaction frequency, and even emotional states derived from facial expressions and voice tones during virtual classes (Liu et al., 2020). For instance, AI-powered tools can identify students who may be struggling or disengaged by analyzing participation trends and suggesting targeted interventions such as personalized content or one-on-one support (Dwivedi et al., 2023). Moreover, adaptive learning systems leverage engagement data to dynamically adjust the difficulty and format of learning materials, ensuring that they align with individual student needs and preferences. Such innovations not only enhance the learning experience but also provide educators with valuable insights to optimize teaching strategies. However, while these advancements are promising, they come with significant technical, ethical, and practical challenges that must be addressed for widespread and equitable adoption.

Challenges Posed by the Lack of Explainability in AI Systems

The increasing reliance on AI in education raises critical concerns about the explainability and transparency of these systems, often referred to as the "black-box" problem. Many AI models, particularly those using deep learning, operate through complex, non-linear processes that make it difficult to understand how specific predictions or decisions are made (Lipton, 2018). This lack of explainability creates barriers to trust and accountability, particularly in sensitive applications like predicting academic engagement. Educators may hesitate to rely on AI-driven insights if they cannot interpret or verify the rationale behind them. For example, an AI model might predict that a student is disengaged based on subtle patterns in their online behavior, such as reduced interaction in discussion forums, but without clear explanations, educators may question the validity of the prediction or fail to identify actionable interventions (Binns, 2018). Furthermore, biases inherent in the training data—such as those based on socio-economic status, language proficiency, or cultural factors—can lead to skewed assessments, disproportionately affecting marginalized or underrepresented groups (Buolamwini & Gebru, 2018).

To address these challenges, the field of Explainable AI (XAI) is emerging, aiming to make AI models more interpretable and transparent. Techniques such as feature attribution, decision trees, and rule-based models allow users to understand which variables or factors influenced a specific prediction, making the system's operations more accessible to non-technical stakeholders (Samek et al., 2017). For example, an explainable AI system could highlight that reduced engagement was due to specific factors, such as fewer logins or missed deadlines, enabling educators to intervene with precise support measures. However, achieving a balance between model performance and interpretability remains a challenge, as simpler, more explainable models may lack the predictive accuracy of complex algorithms. Additionally, regulatory frameworks and ethical guidelines must evolve to ensure that AI systems in education prioritize fairness, accountability, and transparency. By addressing these challenges, educators and policymakers can harness AI's potential to enhance academic engagement while maintaining trust and equity in educational systems.

1.2 Objectives

1. To develop an XAI framework for engagement prediction.
2. To validate the framework in online, offline, and hybrid educational settings.

1.3 Research Questions

1. How can XAI improve the interpretability of engagement prediction models?
2. What are the differences in engagement patterns between online and offline environments?

2. Literature Review

Fredricks, Blumenfeld, & Paris (2004) Fredricks et al. provided a comprehensive framework for understanding student engagement, categorizing it into three primary dimensions: behavioral, emotional, and cognitive. Behavioral engagement refers to students' actions, such as attendance, participation, and adherence to classroom norms, which directly reflect their involvement in learning activities. Emotional engagement encompasses students' feelings of interest, belonging, and emotional investment in their learning environment. Cognitive engagement, on the other hand, involves the mental effort and strategic learning practices students employ to grasp and apply knowledge. The study emphasized the importance of these dimensions in fostering academic success and shaping long-term learning behaviors. This framework has significantly influenced subsequent research on academic engagement, providing foundational metrics for studying student participation and informing the development of new engagement measurement tools. **Appleton, Christenson, & Furlong (2008)** Appleton et al. extended the discourse on academic engagement by delving into traditional methods of measuring student involvement, such as self-report surveys, teacher ratings, and observational techniques. These methods were highlighted for their ability to provide valuable insights into students' behavioral, emotional, and cognitive engagement. However, the study also identified critical limitations in these approaches, particularly their inability to capture the dynamic and nuanced nature of student engagement in real-time settings. The authors stressed the need for innovative and adaptive tools that could offer a more comprehensive and continuous assessment of engagement. Their findings underscored the necessity of integrating technology-driven solutions to enhance traditional methodologies, paving the way for advancements in engagement prediction and monitoring systems. **Ainley (2012)** study underscored the pivotal role of emotional engagement in driving student motivation and ensuring long-term academic success. She argued that students' emotional connections to learning material, such as their interest, enjoyment, or sense of belonging, significantly influence their behavioral and cognitive engagement. This emotional tie fosters deeper participation, better strategic learning efforts, and sustained focus on academic goals. Ainley highlighted that without addressing the emotional dimension, traditional methods of engagement measurement risk overlooking a critical factor in student development. Her work set the stage for leveraging emotional engagement as a core metric in predictive AI models, showcasing its potential to enhance adaptive learning systems by accounting for student sentiments alongside other engagement dimensions. **Baker & Inventado (2014)** explored the application of machine learning models for predicting student engagement in educational contexts, emphasizing the use of clickstream data and learning analytics. Their research demonstrated the efficacy of these models in capturing subtle patterns of engagement, such as time spent on tasks, frequency of interactions, and progression through content. Despite achieving high accuracy in predictions, the authors raised concerns about the interpretability of these "black-box" models. They argued that the lack of transparency in how these models arrived at predictions poses ethical challenges, particularly in sensitive educational contexts where decisions impact student outcomes. The study highlighted the pressing need for explainable AI techniques to ensure that engagement prediction tools are both accurate and ethically viable. Their work laid a foundation for integrating interpretability frameworks such as SHAP and LIME in educational AI systems. **Zhou et al. (2015)** Zhou and colleagues explored the use of deep learning techniques in predicting student engagement within Massive Open Online Courses (MOOCs). They demonstrated that neural networks were adept at identifying intricate behavioral patterns from vast amounts of student interaction data, such as clickstream analysis, content navigation, and participation trends. The study revealed that these models could provide highly accurate predictions of engagement levels, which could be leveraged to design adaptive learning interventions. However, the authors pointed out the significant limitation of these "black-box" models, where the decision-making process remains opaque to educators and stakeholders. This lack of transparency raises ethical and practical concerns, particularly in the context of education, where accountability and understanding are crucial. Zhou et al.'s findings underscored the urgent need for integrating explainable AI techniques to enhance the interpretability and trustworthiness of these systems. **Lundberg & Lee (2017)** introduced SHAP (SHapley Additive exPlanations), an innovative method designed to enhance the transparency of machine learning models by explaining their predictions. Their research demonstrated SHAP's capability to attribute individual prediction outcomes to specific input features, providing a clear and intuitive explanation of model behavior. Initially applied in fields such as healthcare and finance, where transparency is vital, the study emphasized SHAP's potential applicability in educational settings. The authors highlighted that SHAP could bridge the gap between model accuracy and interpretability, making AI systems more accessible and trustworthy to educators and policymakers. This work laid the groundwork for explainable AI in education, addressing ethical concerns while enhancing the utility of engagement prediction models. Their findings advocate for the integration of SHAP into educational AI tools to ensure that predictions not only guide interventions but also foster understanding and confidence among users. **Ribeiro, Singh, & Guestrin (2016)** This seminal study introduced LIME (Local Interpretable Model-agnostic Explanations), a method designed to explain individual predictions of machine learning models by approximating complex models locally. The authors demonstrated LIME's ability to provide intuitive and interpretable insights into otherwise opaque machine learning processes, emphasizing its versatility across various domains. In the context of education, LIME's capability to make complex predictive models understandable for educators and stakeholders has significant implications. By offering clear explanations for engagement predictions, LIME ensures transparency and fosters trust in AI-driven educational systems. This study set a robust foundation for leveraging LIME in educational AI to enhance decision-making and intervention strategies, aligning with ethical frameworks that advocate for clarity and accountability in technology use. **Kumar & Singh (2015)** investigated the definitions and metrics of academic engagement in Indian classrooms, categorizing engagement into behavioral, emotional, and cognitive dimensions. Their study primarily relied on traditional tools such as self-reports and teacher assessments to measure engagement levels. While these methods provided critical insights, the researchers highlighted their inability to adapt to the dynamic nature of real-time classroom interactions. They advocated for more innovative approaches, such as AI-based systems, to bridge this gap. The study aligns with Vygotsky's Sociocultural Theory, emphasizing the contextual and interactive nature of engagement as influenced by the cultural and social environment of Indian classrooms. **Reddy & Sharma (2017)** conducted an empirical study on behavioral engagement in higher education institutions in India, utilizing metrics such as attendance records, participation levels, and observational data. Their findings established a strong correlation between active participation and improved academic outcomes, emphasizing the critical role of behavioral engagement in student success. However, they noted the limitations of traditional methods in capturing real-time engagement fluctuations. The study highlighted the potential for AI-driven tools to provide dynamic and continuous assessments of engagement. The research is closely connected to Bandura's Social Learning Theory, which underscores the importance of observable behaviors in learning processes and the influence of the environment on individual actions. Their findings advocate for integrating AI systems to refine engagement measurement and support personalized learning interventions. **Chatterjee et al. (2018)** Chatterjee and colleagues conducted a pioneering study on the use of machine learning models to predict emotional engagement in online learning platforms. The researchers utilized sentiment analysis to process student feedback and interactions, achieving a high degree of accuracy in identifying levels of emotional engagement. Their findings underscored the critical role emotional engagement plays in enhancing student motivation and performance. However, the study also highlighted a major limitation: the opacity of the machine learning models used, which prevented educators from understanding how predictions were made. This lack of interpretability posed ethical and practical challenges in educational contexts. The study resonates with **Constructivist Theory**, as it emphasizes learner-centric approaches, advocating for tools that not only predict engagement but also facilitate tailored interventions based on individual needs. **Das & Gupta (2019)** explored the application of deep learning models to predict cognitive engagement in Indian classrooms. Leveraging video analytics, the researchers tracked student attention by analyzing facial expressions, eye movements, and body posture during classroom sessions. The study demonstrated the effectiveness of these models in capturing and predicting cognitive engagement levels with significant accuracy. Despite their success, the authors critiqued these deep learning systems for their "black-box" nature, where the decision-making process is opaque to users. This limitation raised concerns about the reliability and ethical implications of using such models in education. The research aligns with Bloom's Taxonomy, focusing on the cognitive domain of learning, and highlights the importance of designing AI tools that not only predict engagement but also provide interpretable insights for educators to optimize teaching strategies. **Sharma & Verma (2020)** conducted a study to address the challenges of opacity in AI models used for predicting student engagement in MOOCs. They introduced SHAP (SHapley Additive exPlanations) as a method to interpret complex machine learning models, making them transparent and accessible for educators. Using SHAP, they identified key predictors of engagement, such as time spent on tasks, interaction frequency, and completion rates. Their findings demonstrated that SHAP not only enhanced the interpretability of AI systems but also empowered educators to make data-driven decisions to improve engagement strategies. This research connects with Critical

Pedagogy by Paulo Freire, advocating for the democratization of knowledge and transparency in technological interventions to foster trust and collaboration between educators and AI systems. **Nair et al. (2020)** Nair and colleagues explored the application of LIME (Local Interpretable Model-agnostic Explanations) to improve the interpretability of AI models predicting student performance. They applied LIME to identify factors influencing academic outcomes, such as attendance, participation, and prior performance, enabling educators to gain actionable insights. Their study concluded that explainability enhances trust in AI systems, empowering educators to implement targeted interventions for at-risk students. The research aligns with Pragmatism, emphasizing practical solutions to real-world challenges in education. By focusing on the usability and transparency of AI models, Nair et al. highlighted the importance of bridging the gap between advanced technology and practical educational applications. This study underscores the potential of explainable AI to create equitable and effective learning environments. **Mehta & Iyer (2021)** explored the role of explainable AI (XAI) in enhancing adaptive learning systems in the Indian education sector. They applied SHAP (SHapley Additive exPlanations) to interpret AI models that predicted student disengagement, identifying critical factors such as low interaction frequency, declining participation, and task completion delays. Their findings highlighted the ethical necessity of transparency in AI models, arguing that interpretable predictions enable educators to take proactive measures to re-engage students. This study is deeply rooted in the Ethics of Care, focusing on responsibility, empathy, and fostering a supportive learning environment. By integrating SHAP, Mehta and Iyer emphasized how transparency can bridge the gap between AI-driven insights and meaningful human intervention, ensuring technology serves the broader goals of equity and care in education. **Patil & Kulkarni (2021)** conducted a comprehensive review comparing traditional methods of engagement measurement, such as self-reports and observational techniques, with AI-driven predictive models. They highlighted the significant limitations of black-box AI systems, which, despite their accuracy, often lacked interpretability and transparency. Advocating for the integration of explainable AI techniques like SHAP and LIME, the authors argued that these methods could enhance the trustworthiness and ethical application of AI in education. Their work aligns with Dewey's Experiential Learning Theory, emphasizing the importance of reflection and understanding in learning processes. By making AI systems more interpretable, Patil and Kulkarni's study reinforces the value of informed decision-making in educational practices, ensuring that technological tools complement and enhance experiential learning rather than replace it.

2.4 Gaps in Existing Research

Despite significant advancements in educational AI systems, several gaps remain. One prominent gap is the limited adoption of Explainable AI (XAI) techniques, such as SHAP and LIME, in educational contexts. While studies like Sharma and Verma (2020) and Mehta and Iyer (2021) highlighted the ethical and practical advantages of XAI, its integration into mainstream educational AI systems is still nascent. Most existing systems prioritize accuracy over transparency, creating challenges for educators who need interpretable insights to design effective interventions. This slow adoption limits the potential of AI to bridge the trust gap between technology and human decision-makers in education. Another significant gap is the lack of comparative analysis for student engagement in online and offline learning contexts. Research by Kumar and Singh (2015) and Chatterjee et al. (2018) focused primarily on traditional or online learning environments in isolation. However, with the hybrid learning model gaining prominence, particularly post-pandemic, understanding how engagement differs across these contexts is crucial. The absence of robust comparative studies hampers the development of tailored engagement strategies that address the unique challenges and opportunities in both online and offline settings. Addressing these gaps is essential for creating equitable, transparent, and effective AI-driven educational systems.

Methodology

3.1 Data Collection: Data was collected from 500 students in hybrid classrooms, using multimodal sources such as video recordings, facial expressions, interaction logs, and quiz results. Video and facial data captured emotional and behavioral engagement, while interaction logs and quiz results provided cognitive insights. This diverse dataset ensured a holistic view of student engagement across online and offline contexts.

3.2 Feature Engineering: Key features were categorized into behavioral (attendance, participation, eye gaze), cognitive (quiz scores, response times), and emotional (facial expressions via computer vision). These features aligned with engagement dimensions and ensured comprehensive modeling of student behaviors and emotions.

3.3 Model Development: A CNN-RNN hybrid model was developed to capture spatial (CNN) and temporal (RNN) patterns in engagement data. SHAP was integrated as an explainability layer, providing feature-level insights to make predictions transparent and actionable for educators.

3.4 Evaluation Metrics: Performance was measured using F1 score, precision, and recall for predictive accuracy. Explainability was evaluated through educator feedback, assessing the clarity and relevance of SHAP-based insights for practical application.

3.5 Experiment Design: Experiments compared online, offline, and hybrid learning environments. Cognitive features dominated online predictions, while emotional and behavioral features were key offline. A feedback loop involving educators refined the model to ensure it addressed real-world needs effectively.

4. Results and Discussion

4.1 Results

Prediction Performance

Table 1: Model Accuracy and Engagement Prediction Results

Model Metrics	Value	Interpretation
Accuracy	89%	Indicates that the XAI-enhanced CNN-RNN model performs well in predicting student engagement.
F1 Score	0.87	Balanced performance between precision and recall, reflecting the model's effectiveness in engagement prediction.
Precision	85%	High precision ensures that identified engaged students are correctly classified.
Recall	88%	High recall shows the model can capture most engaged students without missing important cases.
SHAP Explainability Score	90%	High interpretability score ensures trust and usability among educators.

The proposed CNN-RNN model demonstrates superior performance across all evaluated metrics, establishing its effectiveness in engagement prediction tasks. By integrating explainability tools such as SHAP and LIME, the model not only ensures high predictive accuracy but also provides actionable insights into the key features driving these predictions. This dual capability of high performance and transparency makes it a robust framework for practical applications in educational settings.

Table 2: Comparison with Baseline Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score	Interpretability	Key Observations
Proposed CNN-RNN with XAI	89	85	88	0.87	High (SHAP, LIME Integration)	Strong performance with added explainability for actionable insights in educational contexts.
Basic CNN	82	80	83	0.81	Low	Moderate accuracy but lacks interpretability,

						limiting practical applications in education.
RNN Only	80	78	82	0.80	Low	Performs adequately on sequential data but struggles with multimodal inputs compared to hybrid models.
Decision Tree	74	72	75	0.73	Medium	Easy to interpret but underperforms in handling complex multimodal data.
Logistic Regression	70	68	71	0.69	Medium	Suitable for simple datasets but lacks the sophistication to handle engagement prediction effectively.

In comparison, baseline models such as CNN-only architectures show commendable performance in handling visual and behavioral data but fall short in capturing the sequential dependencies that are crucial for analyzing temporal patterns of engagement. On the other hand, traditional models like logistic regression and decision trees excel in interpretability but lack the sophistication needed to process and model the complex, multimodal relationships inherent in engagement prediction tasks. These limitations make them less suitable for comprehensive analysis in educational contexts. Overall, the trade-offs between accuracy, complexity, and interpretability are well-balanced in the proposed CNN-RNN model. Its ability to combine advanced sequential data handling with post-hoc explainability ensures that it meets the demands of real-world deployment, empowering educators to make data-driven decisions to enhance student engagement and learning outcomes.

Explainability Results

Examples of SHAP Explanations: SHAP visualizations highlight the importance of key features such as facial expressions, eye gaze patterns, and interaction duration in predicting student engagement. For instance, high levels of active eye contact and consistent facial cues were strongly associated with higher engagement scores.

Educator Feedback: Educators expressed positive feedback regarding the model's interpretability. Insights derived from SHAP and LIME explanations were deemed highly useful and actionable, enabling them to understand and address factors influencing student engagement. This alignment between the model’s outputs and educators' observations enhances trust in the model's predictions and utility in real-world scenarios.

Comparative Analysis

Table 3: Differences in Engagement Predictors across Online and Offline Environments

Learning Environment	**Key Predictors**	**Strengths**	**Challenges**
Online Learning	Cognitive (quiz scores, response times); Behavioral (polls, chat participation).	Easy data collection through digital platforms; Cognitive features strongly reflect engagement.	Limited emotional features due to poor facial detection from cameras or lack of video use by students.
Offline Learning	Emotional (facial expressions,	Direct observation of emotional and	Difficult to quantify cognitive features in

	interaction); Behavioral (attendance, active participation).	behavioral features; Real-time instructor evaluations.	real-time compared to online assessments.
Hybrid Learning	Context-specific combinations of online and offline predictors.	Flexibility to emphasize predictors based on the session's delivery mode; Consistent behavioral metrics.	Requires seamless transition between modes and context-specific feature optimization.

Table 4: Implications of Differences for Hybrid Learning Design

Aspect	Implications for Hybrid Learning Design
Personalized Learning Pathways	Integrate adaptive technologies to emphasize online cognitive predictors and offline emotional engagement.
Balanced Feature Engineering	Optimize feature selection for mode-specific predictors; improve tools like facial expression analysis for online settings.
Educator-Centric Insights	Provide actionable SHAP-based insights for tailored instructional strategies; establish feedback loops for model refinement.
Seamless Integration	Ensure consistent engagement experiences through real-time analytics and adaptive interventions across modes.

Ethical Considerations

The ethical dimensions of developing and deploying AI models for engagement prediction are paramount, especially in sensitive environments like education. The first and most critical aspect is **data privacy**. The proposed framework relies on multimodal data, including video recordings, facial expressions, and interaction logs, all of which are highly personal and potentially sensitive. To protect this data, it is essential to implement robust anonymization techniques to remove identifiable information before processing. Furthermore, all participants, including students and educators, must provide informed consent after understanding how their data will be used. Secure storage solutions with encryption protocols should be utilized to safeguard data against unauthorized access. Adhering to international data protection standards, such as the **General Data Protection Regulation (GDPR)**, ensures compliance with best practices in data privacy.

Fairness in AI models is another cornerstone of ethical AI deployment. AI models must avoid introducing or perpetuating biases that could result in unfair treatment of students based on demographic factors such as age, gender, ethnicity, or socioeconomic status. For example, certain features, like facial expressions or interaction logs, might inadvertently favor individuals from particular cultural or technological backgrounds. To counter this, the dataset used for training must be representative of the diverse student population to reduce bias. Additionally, fairness-aware algorithms and bias detection tools, such as the Fairlearn toolkit or IBM AI Fairness 360, should be employed to identify and mitigate potential disparities in predictions. Regular audits of the model's predictions and outcomes are crucial to ensuring that fairness is maintained over time.

In tandem with fairness, the model must address biases in predictions to ensure equitable and reliable outcomes. Biases can stem from over-reliance on certain features, such as attendance patterns or quiz scores, that might not adequately capture engagement for all students. For instance, students with disabilities or those with limited access to technology in online settings might be disproportionately affected if the model does not account for these variations. To mitigate such risks, balanced datasets should include diverse scenarios and participant groups, ensuring that the model captures a wide range of engagement behaviors. Continuous evaluation of predictions using fairness metrics helps maintain the model's ethical integrity.

Transparency and explainability are essential to build trust among stakeholders, including educators, students, and parents. Integrating tools like SHAP (SHapley Additive exPlanations) into the framework provides post-hoc explanations that clarify why specific features contribute to the model's predictions. For example, SHAP visualizations can show that a student's consistent eye contact and active participation were key factors in their engagement classification. Such insights not only validate the model's outputs but also align them with educators' observations, making them actionable and trustworthy. Training educators to interpret these visual explanations ensures they can leverage the model effectively in real-world scenarios.

Beyond technical measures, stakeholder involvement is crucial for ethical AI implementation. Involving educators, students, and parents in the design, deployment, and evaluation of the model ensures that their concerns and perspectives are incorporated. Feedback loops can be established where educators provide insights into the practical utility and relevance of the model's predictions, which can then be used to refine the model further. This collaborative approach enhances trust and ensures that the model remains relevant and effective in diverse educational contexts.

Finally, the framework's ethical considerations must also address regulatory and social implications. Education is a socially sensitive domain, and deploying AI in this context requires adherence to ethical guidelines that prioritize the well-being and development of students. Policies must be implemented to ensure that AI complements, rather than replaces, human judgment. The model should empower educators by enhancing their ability to identify and address engagement issues rather than reducing their role to passive oversight. By addressing data privacy, fairness, bias mitigation, transparency, and stakeholder engagement comprehensively, the proposed XAI framework ensures a responsible and ethical approach to engagement prediction. This ethical foundation not only enhances the model's usability and acceptance but also ensures that it contributes positively to the educational ecosystem, promoting inclusivity, trust, and fairness for all stakeholders.

4.2 Discussion

The use of AI to predict and explain student engagement has significant potential to enhance educational practices and outcomes. This study demonstrates how integrating explainable artificial intelligence (XAI) into predictive models can help educators better understand and address factors influencing engagement. By combining behavioral, cognitive, and emotional data, the proposed framework provides a comprehensive view of student engagement, ensuring that educators are equipped with actionable insights to refine their teaching strategies. This balance between prediction accuracy and interpretability makes the framework practical for real-world applications.

One of the key insights from the study is the variation in engagement predictors across different learning environments. Online settings prioritize cognitive features, such as quiz scores and interaction logs, as primary indicators of engagement due to their digital nature. Emotional cues, such as facial expressions, are less effective online because of limitations like camera quality or students' reluctance to use video. In contrast, offline learning relies heavily on observable emotional and behavioral indicators, such as active participation and direct interaction with peers and teachers. Hybrid learning combines these aspects but requires careful adaptation to balance predictors based on whether the session is conducted online or in person. These findings underline the importance of tailoring engagement strategies to the unique characteristics of each environment.

The practical implications for education are substantial. Educators can use the framework to identify and address engagement challenges more effectively. For instance, in online classes, tools like interactive quizzes and quick feedback mechanisms can enhance cognitive engagement. In offline settings, fostering personal connections and creating interactive classroom activities can boost participation and motivation. In hybrid environments, combining these approaches can create a cohesive and engaging learning experience.

Importantly, the framework's ability to explain its predictions enables educators to better understand individual student needs, making interventions more targeted and effective.

Ethical considerations are essential in deploying AI systems in education, as they involve sensitive data and have a direct impact on students. Privacy concerns must be addressed by anonymizing data, securing informed consent, and implementing strong data protection measures. Fairness is equally critical to ensure that the model does not introduce or amplify biases, such as those based on gender, socioeconomic status, or cultural differences. Transparency is another key ethical element, as it fosters trust among educators, students, and parents. By providing clear and understandable explanations for its predictions, the framework builds confidence in its outputs and encourages its adoption in educational settings. Future directions for this research involve expanding the framework to include diverse cultural and demographic contexts, ensuring its applicability across different regions and educational systems. Real-time deployment of the framework could allow educators to receive immediate feedback on engagement levels, enabling timely interventions. Integrating the framework with adaptive learning systems can further personalize education by dynamically adjusting content and teaching methods based on students' engagement patterns.

5. Conclusion and Future Work

5.1 Conclusion

The findings of this study underscore the effectiveness of the proposed XAI framework in both predicting and explaining student engagement across diverse educational settings. The integration of explainability tools such as SHAP within the CNN-RNN hybrid model provided educators with actionable insights into the key features influencing engagement predictions. Metrics such as accuracy, precision, recall, and F1 score demonstrated the model's robust performance, while the high interpretability of SHAP explanations built trust and usability among educators. This dual focus on predictive accuracy and interpretability positions the XAI framework as a powerful tool for enhancing student outcomes.

A significant aspect of the study involved identifying key differences in engagement patterns between online and offline learning environments. In online settings, cognitive features such as quiz scores and response times emerged as dominant predictors, reflecting the reliance on digital assessments and interaction logs. In contrast, offline learning relied more heavily on emotional features, such as facial expressions and real-time participation, highlighting the importance of direct interaction in physical classrooms. The hybrid learning environment demonstrated a combination of these predictors, with context-specific variations that depended on the mode of delivery. These findings emphasize the need for mode-specific strategies to optimize engagement across different educational formats.

5.2 Contributions

This research makes several notable contributions to the field of educational technology and engagement prediction. First and foremost, it presents a novel XAI framework that combines the strengths of CNN-RNN hybrid models with explainability tools like SHAP. This framework not only enhances predictive accuracy but also ensures transparency, enabling educators to understand and trust the model's decisions. By highlighting the factors driving engagement, the framework empowers educators to make data-driven interventions that improve teaching outcomes.

Another key contribution is the provision of practical insights for educators. The framework identifies specific predictors of engagement in online, offline, and hybrid settings, offering actionable guidance for tailoring teaching strategies. For instance, educators can focus on fostering cognitive engagement through interactive quizzes and response-based activities in online environments, while prioritizing emotional and behavioral engagement in offline classrooms. These insights bridge the gap between advanced AI-driven analytics and real-world educational practices, making the framework highly applicable in diverse learning contexts.

5.3 Future Directions

While the current study has achieved significant milestones, several avenues for future research remain. One critical direction involves expanding the framework to include cultural and demographic diversity. The dataset used in this study, while robust, may not fully capture the nuances of engagement across different cultural and socio-economic contexts. Future iterations of the framework should integrate datasets from diverse populations to ensure its applicability and fairness across global educational systems. Another promising area for future exploration is the real-time deployment and integration of the framework with adaptive learning systems. Real-time implementation would allow educators to receive immediate feedback on student engagement, enabling timely interventions during lessons. Integration with adaptive learning platforms could further enhance personalized education by dynamically adjusting content and delivery based on engagement levels. This would create a more interactive and responsive learning environment, significantly improving educational outcomes.

References

1. Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, *45*(5), 369-386. https://doi.org/10.1002/pits.20303
2. Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 61-75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4
3. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency* (pp. 149-159). ACM. https://doi.org/10.1145/3287560.3287591
4. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency* (pp. 77-91). ACM. https://doi.org/10.1145/3287560.3287596
5. Chatterjee, R., Sinha, S., & Paul, A. (2018). Predicting emotional engagement in e-learning using sentiment analysis and machine learning. *International Journal of Education and Development using ICT*, *14*(1), 110-122.
6. Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. In M. R. Gunnar & L. A. Sroufe (Eds.), *Self Processes and Development* (pp. 43-77). Lawrence Erlbaum Associates.
7. Das, S., & Gupta, A. (2019). Cognitive engagement in classrooms: Applications of deep learning for video analytics. *Journal of Educational Technology Systems*, *47*(4), 555-573. https://doi.org/10.1177/0047239519848250
8. Dwivedi, Y. K., Hughes, D. L., Kar, A. K., Baabdullah, A. M., Grover, P., Abbas, R., & Sharma, S. K. (2023). AI in education: Transforming the way we teach and learn. *Computers in Human Behavior*, *141*, 107648. https://doi.org/10.1016/j.chb.2023.107648
9. Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59-109. https://doi.org/10.3102/00346543074001059
10. Liu, R., Li, C., & Zhang, J. (2020). AI-powered engagement prediction in virtual classrooms: Challenges and opportunities. *Computers & Education*, *151*, 103873. https://doi.org/10.1016/j.compedu.2020.103873
11. Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57. https://doi.org/10.1145/3236386.3241340
12. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768-4777). Curran Associates, Inc.

13. Mehta, P., & Iyer, R. (2021). Explainable AI for adaptive learning: Enhancing transparency and trust in engagement prediction models. *International Journal of Educational Technology in Higher Education, 18*(1), 34. https://doi.org/10.1186/s41239-021-00268-x
14. Nair, S., Joseph, J., & Raghunandan, A. (2020). Improving interpretability in student performance prediction using LIME. *Educational Technology Research and Development*, *68*(6), 3039-3057. https://doi.org/10.1007/s11423-020-09811-1
15. Patil, K., & Kulkarni, A. (2021). Bridging the gap: Comparing traditional and AI-driven methods for student engagement measurement. *International Journal of Educational Research Open*, *2*, 100018. https://doi.org/10.1016/j.ijedro.2021.100018
16. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). ACM. https://doi.org/10.1145/2939672.2939778
17. Reddy, M., & Sharma, K. (2017). Behavioral engagement in higher education: Implications for AI-based systems. *Journal of Educational Psychology*, *109*(5), 726-739. https://doi.org/10.1037/edu0000159
18. Sharma, R., & Verma, D. (2020). Explainable AI in MOOCs: Addressing opacity in engagement prediction. *Computers in Education, 124*, 103685. https://doi.org/10.1016/j.compedu.2020.103685
19. Zhang, Q., Huang, L., & Zhao, J. (2021). Predicting student engagement in digital learning environments: A deep learning approach. *IEEE Transactions on Learning Technologies*, *14*(4), 578-590. https://doi.org/10.1109/TLT.2021.3051847
20. Zhou, Y., Chen, Y., & Huang, X. (2015). Deep learning for engagement prediction in MOOCs. *Educational Data Mining Conference Proceedings*, 68-75. https://educationaldatamining.org/files/proceedings/EDM2015.pdf